The Thesis Committee for Emily Camille Pease
certifies that this is the approved version of the following thesis:

Theory-Guided Data Science: Combining Machine Learning with Domain Expertise to Predict Springflow

Committee:

Suzanne A. Pierce, Supervisor

Michael Pyrcz

Yolanda Gil

Theory-Guided Data Science: Combining Machine Learning with Domain Expertise to Predict Springflow

by

Emily Camille **Pease**

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Energy and Earth Resources

The University of Texas at Austin

May 2020

I dedicate this thesis to my loving family , who have supported me no matter what path I chose.

Acknowledgments

I deeply appreciate the support and encouragement from my supervisor, Suzanne A. Pierce. She has helped me navigate the world as a geology-focused data scientist and pushed me to always seek new, exciting opportunities. I am also so inspired by my conversations with Michael Pyrcz, who helped me explore the many capabilities of machine learning and who produces students who must always remember to be a domain expert first. I'm very appreciative of the advice from Yolanda Gil, a fantastic role model as a woman in data science. Additionally, Adam Papendieck was very helpful in the organization of my written thesis work.

There were unfortunately many voices that discouraged me along the way because I took a different path. Despite the hurdles, I persevered through the patriarchy to become a well-versed Pythonista. I am thankful to my support network that helped me navigate these challenging times.

Finally, I am so incredibly grateful to my parents, Frank and Paula Pease, for their continuous support through undergrad and grad school. Additionally, Ross Kushnereit was a huge help in navigating my graduate school experience and preparing for a career. My dog, Lady Bird, was also helpful in providing snuggles after a long day of work or research. I love you all so much.

Emily C. Pease

The University of Texas at Austin

May 2020

Theory-Guided Data Science: Combining Machine Learning with Domain Expertise to Predict Springflow

Emily Camille Pease, M.S.
The University of Texas at Austin, 2020

Supervisor: Suzanne A. Pierce

Traditionally, science follows a theory-based approach through which physical equations are used to model natural phenomena. In this recent era of artificial intelligence and "big data", there is a shift into a new paradigm of scientific discovery. The paradigm of theory-guided data science (TGDS) enables scientists to perform data science modeling while retaining their domain expertise to produce informed results consistent with the physical system. Predicting springflow discharge from Comal Springs using machine learning was determined to be an appropriate case study. The Edwards Aquifer in central Texas is one of the largest aquifers in the world and serves as the primary water supply for over 1.5 million Texans, providing water for recreational activities, businesses, and down-stream users. Additionally, these waters serve as a home to many aquatic species, eight of which are endangered or threatened. Quantifying springflow is essential in regulating groundwater resources

in the Edwards Aquifer, especially during drought conditions. Here, a theory-guided predictive machine learning model for springflow estimation at Comal Springs is developed. First, feature engineering is performed to discover relations between data available in the Edwards Aquifer region, selected through theory-guided parameter initialization. Next, multiple machine learning models were explored and tested in their ability to model a complex springs system. Finally, theory-guided refinement of data science outputs was performed to make the model results consistent with what is possible in nature.

Contents

List of Tables

List of Figures

Chapter 1

A New Paradigm in Scientific Discovery: Theory-Guided Data Science

1.1 Intelligent Systems and Geosciences

The field of artificial intelligence (AI) includes machine learning, natural language processing, image recognition, robotics, computer vision, and more. AI enables scientists to quickly process large, multidimensional datasets to discover underlying patterns and visualize results. In this era of "big data", new data is generated constantly, often without having a pre-developed scientific hypothesis.

The geosciences span a broad set of subdisciplines across the polar, atmospheric, geospace, subsurface, ocean, and deep earth studies. Intelligent Systems and Geosciences (IS-GEO) is a movement to foster a community across disciplines, identify barriers in research, increase communication and encourage long-term inter-

disciplinary collaborations. Understanding results of research from the geosciences and accelerating the time from discovery to use in societally relevant decisions is a key challenge for modern times [12]. Already the pace of geoscientific research has accelerated, thanks to the use of advanced computing capabilities. Yet, the majority of computing and cyberinfrastructure collaborations have come from non-AI research [12]. In order to achieve increasing returns from interdisciplinary AI and geoscientific research, approaches need to expand to include techniques from AI that can assist geoscientist with reasoning about their data and provide Earth-based insights to inform the application of AI to complex problems.

It is essential for scientists to be able to access, store, and analyze the continuous stream of data in order to inform decision making. The research presented here explores strategies for combining approaches from artificial intelligence with geoscientific understanding of groundwater problems. Specifically, this research applies machine learning methods to predict springflow discharge. Implementation of the various methodologies requires the use of statistical machine learning approaches together with knowledge of groundwater and surface water relationships to define features that represent both the geologic expectations of behavior in physical systems, as well as adhering to acceptable machine learning (ML) techniques.

1.2 Domain Centered Approaches to Geoscientific Inquiry, Data, and Models

The geosciences is a broad field, including hydrology, volcanology, sedimentology, ecology, physical geography, and others with each subfield collecting unique types of data to use for analysis, e.g. seismic, well logs, aquifer levels, field measurements, etc. This data is used as inputs in geological models to simulate a simplified natural

system. The type of modeling performed ranges from simple statistical models to finite difference modeling with complex geometries and boundary conditions to represent Earth processes.

The problem formulation in the geosciences has historically been deductive, following the scientific method. A geologist starts with a theory, forms a hypothesis relevant to that theory, then performs quantitative research to confirm or refute the hypothesis. More recently as AI and ML have become more popular across disciplines and science has become fueled by increasing availability of large data and processing, the field of geology has shifted to include inductive reasoning, an approach that examines an already-existing dataset in an attempt to gain new knowledge.

1.3 Combining Machine Learning Approaches with Geoscientific Knowledge

Machine learning is the sub-field of AI through which humans program machines to learn from large amounts of data without human intervention. Methods include regressions, classifications, dimensionality reduction, bootstrap, and more. Machine learning allows data scientists to create models that learn and improve with more data without being trained to solve any specific task. The field of AI has evolved from a field in academic research to an impactful part of everyday life. ML methods can be used to make predictions as well as to gain an understanding of the data, analyze patterns, and perform feature ranking. It has been used in boosting the quality of life by detecting illness, increasing cyber security, transforming business innovation, and accelerating scientific discovery. In the geosciences, data science methods have been used to process large amounts of data quickly to understand natural phenomenon, discover patterns, and make predictions [8][11][13]. In remote

sensing, ML allows for efficient image analysis, classification, and geological mapping [1]. In hydrology, AI has been used to model lake surface temperatures [18], make predictions on streamflow discharge [27], and modeling [19][33]. These are just a few examples of the many uses of ML in the geosciences.

1.4 Theory-Based Modeling

The relationship between two variables in nature is often simulated based on scientific knowledge of physical laws or theories. Historically, scientific knowledge has grown by forming a hypothesis and subsequently testing it by gathering data to test or refute it. This form of scientific discovery is referred to here as theory-based modeling. Theory-based modeling is process-centric, meaning the physical process is simulated using equations and estimated input values; it is driven by prior knowledge of the analytical expressions that govern the physical process. The paradigm before it, experimentation, lacked the ability to handle large datasets and focused on observation to conduct research.

Theory-based modeling comes with limitations. Natural phenomena are often complex with non-stationary patterns. Simplifying assumptions combat natural complexities, but can lead to poor model performance.

1.5 Data Science Modeling

Data science modeling is a data-centric method used by scientists and researchers to understand natural processes through statistical modeling. In contrast to theory-based modeling, data science modeling is often done without a pre-described hypothesis or theory to test. These models instead seek to discover patterns that were

previously undetected and not actively sought out by researchers.

Data science modeling is powerful but many limitations exist, especially when being performed in the geosciences. Scientists are often studying complex processes, and data science model outputs may not align with what is known about the physics in a complex, natural system. Data science only captures associative relationships among features and tells us little about causal relationships in natural phenomena. Another limitation of statistical learning methods is the lack of sufficient data, which is often the case in the geosciences. Machine learning methods, as used in this thesis, require a large subset of data that fairly represents the population in order to make accurate predictions. When working in the petroleum industry, the data scientist is provided a pre-curated dataset containing data their company has collected with no regard for even sampling or coverage. Another problem that presents itself is data reliability. Any biases in the data must be acknowledged and corrected prior to statistical learning. Otherwise, the machine learning model will produce biased outputs.

1.6 Theory-Guided Data Science

Theory-guided data science (TGDS) is a new paradigm of scientific discovery from data that takes advantage of data science methods without ignoring scientific theories and domain expertise [16]. Many researchers are pushing for a synergy between theory-based and data-driven modeling [3][10][12][34]. Traditional "black-box" machine learning can limit the quality of a predictive model in the sciences by focusing only on the statistics and completely ignoring natural processes. TGDS integrates domain expertise and simulators with data science to enhance the quality of model outputs.

Combining theory-based models and data science models in theory-guided data science expands the potential of scientific discovery [16]. Theory-based models often make assumptions to simplify complex phenomena, which can lower model performance. Similarly, data science methods do not account for natural processes and can fail to represent the relationship in scientific problems. A synergy between data science and domain expertise enables knowledge discovery in scientific problems.

Chapter 2

Defining a Geoscientific Context and Problem Space to Apply Theory-Guided Approaches

2.1 Historical Observations and Groundwater Hydrology in the Edwards Aquifer of Texas

2.1.1 Introduction: Karst Aquifers in the United States

Accounting for 20-40% of the US groundwater supply, karst aquifers provide water to millions of people [7]. These aquifers form in soluble rocks, such as limestone or dolomite, and due to their high porosity and permeability, they are often described as "underground rivers" by non-specialists. While this layman's classification is true to an extent, it certainly over-simplifies their complex behavior. Karst aquifers are the result of thousands to millions of years of groundwater flow that dissolved calcium carbonate rock along fractures, causing them to grow and expand into caverns and

regional flow paths [22][35]. They often experience extremely high transmissivities along conduits on the order of 5-10 magnitudes larger than what would be typical of alluvial aquifers, with values as high as 2000 ft^2/day. One issue in studying karst aquifers is that these conduits are difficult to locate. The complexities of karst aquifers extend to surface water-groundwater interactions. Even the use of a range of geophysical surveys of the subsurface is sometimes not adequate in gaining a complete understanding of their behavior [14].

2.1.2 Edwards Aquifer in Central Texas

The Edwards Aquifer in central Texas is one of the largest karst aquifer systems in the United States and serves as the primary water supply for Bexar, Comal, Hays, Medina, and Uvalde counties (over 1.5 million Texans) [36]. The aquifer is comprised of several components (Figure 4.2). The northern most segment is located on the Edwards Plateau and is referred to as the drainage zone. Water that enters this region through precipitation or streamflow can infiltrate into the unconfined aquifer as recharge [22][29], or gets carried further downstream to the recharge zone. The recharge zone is where a majority of the recharge into the Edwards Aquifer occurs and lies along the Balcones Fault Zone. In this component of the aquifer, streams lose flow directly into this unconfined portion of the aquifer [21][22] and during storms, precipitation infiltrates directly into the aquifer through runoff. South of the recharge zone is the artesion (confined) region of the Edwards Aquifer. Water that is carried into or falls onto this region will not infiltrate into the Edwards Aquifer. There are many ways to estimate recharge [20][29][31] in order to understand total aquifer storage, but other proxies exist that are perhaps more telling of Edwards Aquifer aquifer water levels such as the Bexar County

index well J-17, Uvalde County index well J-27, and springflow from Comal Springs. Monitoring water levels are essential to the Edwards Aquifer Authority in order to regulate groundwater withdrawal.

2.1.3 Comal Springs

Comal Springs is the largest spring complex in the southwest, located in New Braunfels, Texas. These springs form the headwaters of the Comal River before discharging into the Guadalupe River, where its waters are then used by millions of downstream users. Additionally, the spring waters serve as one of the most popular recreational areas in Texas, where the are used for rafting, canoeing, swimming, and tubing, bringing thousands of Texans together during the hot, summer months. Many endangered species and flora inhabit these waters that depend on these springs for their existence, making it legally essential to ensure the springs do not go dry due to the Endangered Species Act and the Edwards Aquifer Habitat Conservation Plan (see Section 2.1.6). The species inhabit the waters at the spring orifice and downstream, but also deep inside the karst features from which the water discharges. During times of drought, conserving these spring waters are made a priority in order to protect the endangered species. Through the drought of record, springflow in the region had slowly been decreasing until finally in 1956, Comal Springs levels were so low that its discharge was recorded as 0cfs for the first time in recorded history and remained this low for four and a half months [37]. When the rains came towards the end of 1956 and the Central Texas began recovering from the drought, Comal Springs began flowing again and levels increased from 0cfs to over 250cfs in just under four months. This served as a wake up call to many lawmakers and stakeholders that these springs may not exist in 100 years if pumping continued

without change.

2.1.4 Endangered Species in a Diverse Environment

The Edwards Aquifer ecosystem is possibly the most diverse groundwater ecosystem in the world [36]. The major artesian springs of the Edwards Aquifer provide water for recreational activities, businesses, downstream users [39], and, more importantly from a policy standpoint, provide a home to over forty aquatic species [4], eight of which are endangered or threatened according to the World Wildlife Fund, including the San Marcos Salamander (*Eurycea nana*), Texas Blind Salamander (*Typhlomolge rathbuni*, Figure 2.1), San Marcos Gambusia (*Gambusia georgei*), Fountain Darter (*Etheostoma fonticola*), Peck's Cave Amphipod (*Stygobromus pecki*), Comal Springs Riffle Beetle (*Heterelmis comalensis*), and the Comal Springs Dryopid Beetle (*Stygoparnus comalensis*). The main problem for these species is reductions in springflow caused by increased pumping, urbanization, and poor water quality [4].

2.1.5 Governance of the Edwards Aquifer

The Endangered Species Act (ESA) was established in 1973 and provides legal protection and conservation of endangered and threatened species and their habitats both in the U.S. and internationally. States are provided financial assistance to create conservation programs for listed species. In the early 1990s, a federal lawsuit was brought against the U.S. Fish and Wildlife Service by the Sierra Club on behalf of the endangered species that reside in Comal and San Marcos Springs, two of the largest springs in Texas that discharge from the Edwards Aquifer that provide large quantities of water for downstream users. The lawsuit aimed to ensure that

minimum springflow discharge values from Comal and San Marcos Springs were established and enforced in order to protect the species and their habitat. In 1993, the Court ruled that springflow must be maintained and the Texas Water Commission must submit a plan to assure that Comal and San Marcos Springs do not drop below jeopardy levels. It was then in the hands of the U.S. Fish and Wildlife Service to determine thresholds for "take" and "jeopardy" by mid-March 1993. In order to protect the endangered species and fauna that inhabit the springs, the Edwards Aquifer Authority was created under Senate Bill 1477, which became active on September 1st, 1993. The Edwards Aquifer is not controlled by a Groundwater Conservation District, but is regulated by the Edwards Aquifer Authority (EAA). In order to protect the species, the legislature gave the EAA the authority to limit groundwater withdrawal in order to preserve springflow discharge. As part of the establishment of the EAA, a maximum withdrawal of 450,000 acre-feet was put into law that is purposed for agricultural, domestic, industrial, municipal, and recreational uses, as well as a portion of stream flow in multiple rivers that discharge into the Gulf of Mexico (S.B. 1477 [1993]). In 1995, this was challenged by the Medina County Groundwater Conservation District and others, stating that the establishment of the Edwards Aquifer Authority was unconstitutional because their groundwater regulation powers violated landowners' pumping rights under the rule of capture. However, this was case dismissed by the Texas Supreme Court and the EAA was fully operational on June 28, 1996.

2.1.6 The Edwards Aquifer Habitat Conservation Plan

The Edwards Aquifer Habitat Conservation Plan (HCP) is "intended to support the issuance of an Incidental Take Permit (ITP) which would allow the 'incidental

take' of threatened or endangered species resulting from the otherwise lawful activities involving regulating and pumping of groundwater from the Edwards Aquifer (Aquifer) within the boundaries of the EAA for beneficial use for irrigation, industrial, municipal and domestic and livestock uses, and the use of the Comal and San Marcos spring and river systems for recreational and other activities." This Edwards Aquifer HCP serves the purpose to "not appreciably reduce the likelihood of the survival and recovery of covered species associated with the Aquifer and Comal and San Marcos springs and rivers ecosystems" [6]. In other words, the Edwards Aquifer HCP is intended to secure the survival of covered species that inhabit the waters of Comal and San Marcos Springs, despite lawful groundwater usage in the region.

The HCP is separated into three major categories to ensure all aspects of the ecosystem are conserved. These include habitat protection measures, flow protection measures, and supporting measures. Habitat protection measures are in place to protect the ecosystems surrounding the springs such as old channel restoration, non-native animal species control, and aquatic vegetation restoration around the springs. The flow protection measures include Voluntary Irrigation Suspension Program Option (VISPO), Regional Water Conservation Program (RWCP), Stage V Critical Period Management, and San Antonio Water System (SAWS) Aquifer Storage and Recovery (ASR) [6]. The VISPO is a volunteer program through which eligible holders of water rights can suspend all use or a portion of their allotment in return for compensation. The goal of this program is to preserve 40,000 acre-feet of permitted water that will remain unused in the case of severe drought. The RWCP plan offers incentives to municipalities to encourage conservation in exchange for leaving water unpumped in the aquifer for 15 years. This preserves water for times

of drought when springflow levels are low. The implementation of Stage V Critical Period was added on to the already existing Stages I-IV to reduce groundwater permits by 44% in the case of severe drops in water level. The drought stages are based off of Comal Springs discharge and index well J-17 levels (Figure 2.2). These stages enforce the restriction of groundwater withdrawal by up to 44%, yet large periods of time are spent in drought stages (Figure 2.4 and 2.3). Finally, SAWS ASR is intended to minimize the impacts of long-term drought by transporting groundwater from the Edwards Aquifer during healthy, wet periods to another aquifer to be used when Edwards Aquifer water levels are low. Groundwater systems are a critical geologic resource for the state of Texas. In fact, the state has over 9 major and 22 minor aquifer systems that provide water supplies, support ecosystems, and provide necessary source water to sustain the economic sectors across the state [28]. Considering the importance of groundwater to the state for a myriad of reasons, this study focused in on a key aquifer in the region called The Edwards Aquifer to further refine and apply machine learning methods.

The Edwards Aquifer system, stands out because it is one of the largest aquifers in the world, provides water supplies to San Antonio which is the 9th largest city in the nation, and provides the habitat for key endangered species. Additionally, the Edwards Aquifer is vulnerable to fluctuating climate conditions, ranging from drought to flooding, and particularly sensitive to water quality risks. The following sections highlight key characteristics of this important geologic resource, as well as exploring the key vulnerabilities and behaviors that define the system.

2.2 Drought in Texas

2.2.1 The Drought of Record (1950s)

The drought of record lasted from 1950-1957 and caused widespread damage and distraught across Texas and many southwestern states. Approximately three-quarters of the country was impacted by this seven year drought, ranging regionally from mild to severe [24]. Based on statistical studies of long-term precipitation records, an equivalent drought in the same region has a recurrence interval of 140 years [24]. Texas had not experienced a drought this devastating in recent history. Before 1950, Texans used water without consequences. Large amounts of rainfall in the 1940s led to increased western expansion with few problems associated with water supply [24]. In the early 1950s, a period of drought brought many financial and personal hardships. Precipitation deficiencies ranged from 25 to 225% depending on the drought-affected region. From 1950 to 1960, the number of farms and ranches shrank from 345,000 to 247,000, and the state's rural people declined from more than a third of the population to a quarter [2]. Livestock sale prices plummeted while feed prices increased. The overgrazing of fields left the landscape barren, allowing for mesquite and cedar to intrude. Wealthier ranchers would ship their cattle north to greener fields out of state, and poorer ones sold what they had for very little or spent large amounts in the effort to keep the animals alive [2]. The drought of the 1950s led many ranchers to sell everything and move to urban centers, and this rural to urban movement is still in place today.

2.2.2 Policy Interventions After the Drought of Record

After the drought of record, the Texas Legislature took many steps to ensure that Texas would be better equipped for future droughts. The Texas Water Development

Board (TWDB) was established in 1957 to predict water supply needs across the state and provide funding for water conservation projects. The Legislature issued $200 million to the TWDB to make loans to municipalities to build better water infrastructure and more reservoirs [15]. They also passed the Water Planning Act of 1957, which authorized and instructed the Texas Water Commision (then named the Texas Board of Water Engineers) to develop a plan to meet Texas' future water needs and demands [15]. Between 1957 and 1980, 126 reservoirs were built.

The 1990s drought was a wake-up call for what the future could hold in terms of water shortages and limited availability, driving Texas lawmakers to take preventative measures [15]. The Texas Legislature passed Senate Bill 1 in 1997, which acted to implement a huge change in water planning. Through Senate Bill 1, 16 regional water planning groups were established that included representatives from a variety of industries to provide input for future planning. The TWDB was also required to publish a state water plan every five years.

2.2.3 Drought in the 21st Century: 2011-2014

In 2011, Texas had experienced one of the driest winters on record and that dryness continued through the summer and spring, likely as a result of La Nina beginning in 2010 [15]. The record for the driest 12 consecutive months was broken in 2011, and the previous record for driest 12 months was set in 1956 during the Texas drought of record. The hottest statewide average temperatures for June, July and August were all in 2011. The combined June through August temperature was one of the hottest ever for any state, breaking a record set by Oklahoma during the Dust Bowl. The drought grew worse when the reservoirs and stream levels declined to near record heights, and 2011 did not bring more water through precipitation to the state. This

stressed power plant operations that rely on a constant water supply, the agriculture industry, and consumers buying products that increased in value [15].

The drought greatly impacted the agricultural industry. According to Dr. Travis Miller, professor and Texas AgriLife Extension Service program leader for soil and crop sciences at Texas A&M, crop losses were estimated at $5.2 billion for Texas; 52% of cotton acres produced no crop yield in 2011 [15]. Half of all agricultural income comes from crops and the other half comes from livestock. The livestock feed on hay and when the farmers are unable to grow the hay themselves, such as during drought periods, they pay double or triple the price to import it in from the mid-west. In contrast to the 2011-2014 drought when there were no rains, the 1950s drought of record brought some precipitation - enough to grow hay and for livestock to continue grazing [15].

2.2.4 Climate Shift in Texas: The 100th Meridian

The 100th meridian divides the dry, arid, western half of the United States from the humid, eastern half of the country. This bisection is prevalent through a difference in irrigation techniques (dry land vs. irrigated), regional hydrology, and climate [32]. In recent years, this aridity delineation represented by the 100th meridian has swept east across the United States, which could be especially disastrous in Central Texas, located just east of the 100th meridian. Between the 1850s when John Wesley Powell first conceptualized the bisection of the United States and today, the representation of the 100th meridian has shifted 6^o of latitude to the east [32].

Because Central Texas lies directly east of the 100th meridian, this shift is of great interest to current researchers. The population of Texas is predicted to eclipse 50 million by the year 2050, a 70% increase from today's population [38].

As water demands increase into the future, it is unknown how Texas will adapt to these needs, especially as the bisecting line between wet and dry moves east into or through Central Texas. As the meridian shifts east through the central part of the state, it is feared that changes will be seen in our natural springs and regional hydrology. The region of Central Texas experiences severe drought interspersed with flooding events, making any change in climate potentially disastrous to the natural springs and greenery Texans have loved for generations. The outdoor activities along the rivers and hiking trails is a large part of the culture in Central Texas. It will be unspeakable if this drying climate shift alters or destroys these ecosystems.

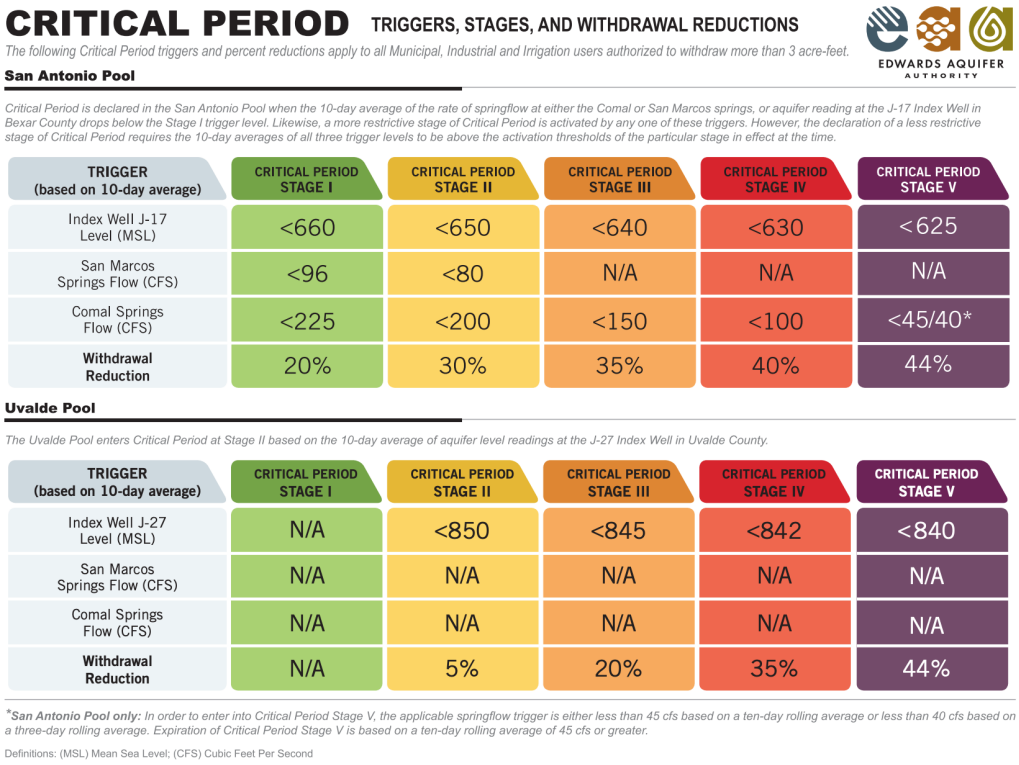
2.2.5 Case study selection and key criteria

Having evaluated the characteristics of the Edwards Aquifer from the perspective of groundwater conditions, risks, vulnerabilities, and governance systems resulted in selection of Comal Springs as a focal point for applying theory-guided approaches. In particular, the evaluation surfaced concerns and key indicators in the following categories: (1) drought, (2) habitat conservation, and (3) water management of the aquifer. Comal Springs is a complex system, yet here the problem is simplified using statistical learning methods. Because the springs are federally protected, there is extensive data covering the region that allows for statistical learning methods to be possible. There are well level and springflow thresholds that act as triggers for critical period stages in the San Antonio section and Uvalde section of the Edwards Aquifer with daily data available, allowing for daily time steps dating to the 1950s (Figure 2.2). When applying machine learning to this case study, these triggers are used to evaluate the accuracy of model outputs.

Comal Springs were chosen because of the amount of data available. It

would be interesting to apply these methods to another springs system to expand the applicability of machine learning for springflow estimation.

Figure 2.1: The Texas Blind Salamander, only present in San Marcos, Texas.



CRITICAL PERIOD TRIGGERS, STAGES, AND WITHDRAWAL REDUCTIONS

EDWARDS AQUIFER AUTHORITY

The following Critical Period triggers and percent reductions apply to all Municipal, Industrial and Irrigation users authorized to withdraw more than 3 acre-feet.

San Antonio Pool

Critical Period is declared in the San Antonio Pool when the 10-day average of the rate of springflow at either the Comal or San Marcos springs, or aquifer reading at the J-17 Index Well in Bexar County drops below the Stage I trigger level. Likewise, a more restrictive stage of Critical Period is activated by any one of these triggers. However, the declaration of a less restrictive stage of Critical Period requires the 10-day averages of all three trigger levels to be above the activation thresholds of the particular stage in effect at the time.

TRIGGER (based on 10-day average)	CRITICAL PERIOD STAGE I	CRITICAL PERIOD STAGE II	CRITICAL PERIOD STAGE III	CRITICAL PERIOD STAGE IV	CRITICAL PERIOD STAGE V
Index Well J-17 Level (MSL)	<660	<650	<640	<630	<625
San Marcos Springs Flow (CFS)	<96	<80	N/A	N/A	N/A
Comal Springs Flow (CFS)	<225	<200	<150	<100	<45/40*
Withdrawal Reduction	20%	30%	35%	40%	44%

Uvalde Pool

The Uvalde Pool enters Critical Period at Stage II based on the 10-day average of aquifer level readings at the J-27 Index Well in Uvalde County.

TRIGGER (based on 10-day average)	CRITICAL PERIOD STAGE I	CRITICAL PERIOD STAGE II	CRITICAL PERIOD STAGE III	CRITICAL PERIOD STAGE IV	CRITICAL PERIOD STAGE V
Index Well J-27 Level (MSL)	N/A	<850	<845	<842	<840
San Marcos Springs Flow (CFS)	N/A	N/A	N/A	N/A	N/A
Comal Springs Flow (CFS)	N/A	N/A	N/A	N/A	N/A
Withdrawal Reduction	N/A	5%	20%	35%	44%

**San Antonio Pool only:* *In order to enter into Critical Period Stage V, the applicable springflow trigger is either less than 45 cfs based on a ten-day rolling average or less than 40 cfs based on a three-day rolling average. Expiration of Critical Period Stage V is based on a ten-day rolling average of 45 cfs or greater.*

Definitions: (MSL) Mean Sea Level; (CFS) Cubic Feet Per Second

Figure 2.2: Critical periods established by the Edwards Aquifer Authority that dictate withdrawal reduction in percentages. Thresholds are set by well levels J-17 and J-27, San Marcos Springs, and Comal Springs.

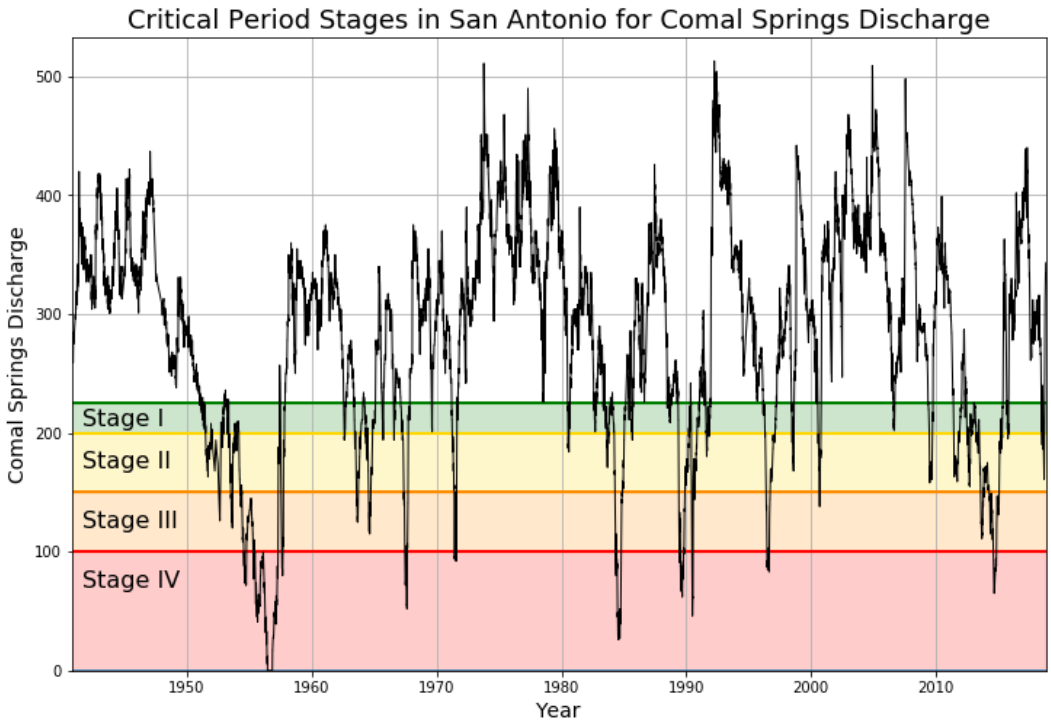


Figure 2.3: Critical periods of the Edwards Aquifer compared with Comal Springs discharge rates.

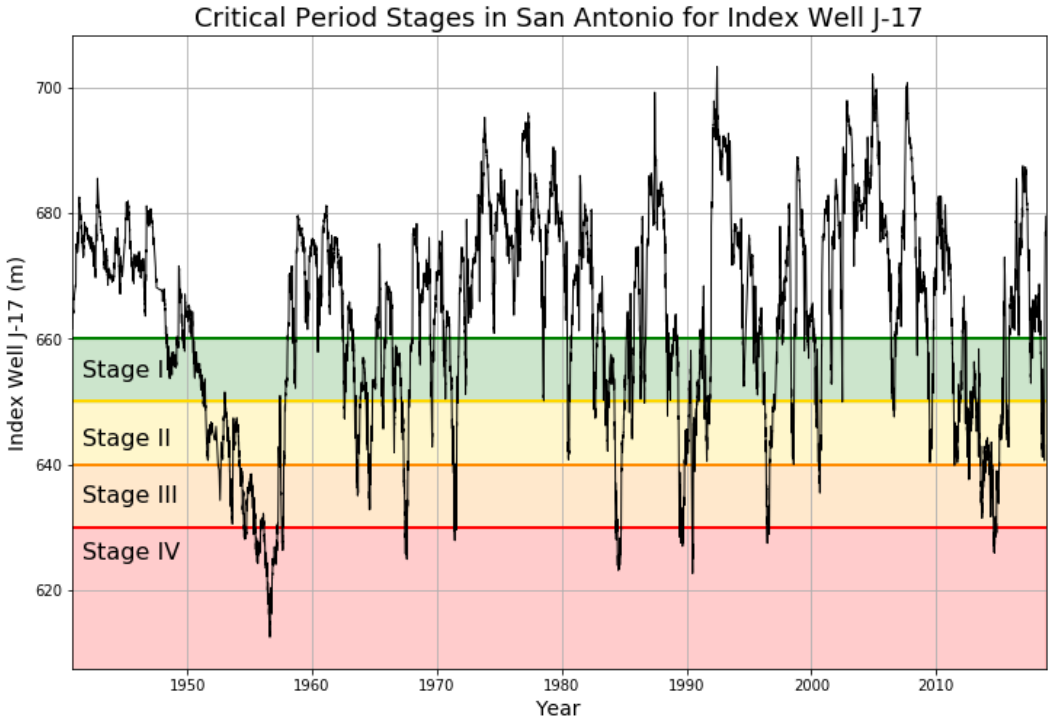


Figure 2.4: Critical periods of the Edwards Aquifer compared with index well J-17 values.

Chapter 3

Defining predictive features for a groundwater system: Comal Springs Case Study

3.1 Theory-Guided Data Science in Predicting Springflow at Comal Springs

3.1.1 Problem

The Edwards Aquifer Authority (EAA) and all stakeholders need accurate and current springflow data in order to make informed decisions for Comal Springs, especially during drought periods when the endangered species that inhabit the springs are most at risk due to low water levels [4]. However, springflow is difficult to measure directly due to the complexity of numerous seeps and spring orifices and the intermittent presence of rainfall and runoff. The U.S. Geological Survey (USGS) is the federal agency responsible for correctly measuring the daily spring discharge

and the EAA is responsible for regulating discharge through groundwater pumping or withdrawal to ensure the springs are flowing above drought thresholds.

The current method used by the USGS for measuring springflow at Comal Springs falls within the third paradigm of scientific discovery, numerical simulation, and involves separating the springflow component from the streamflow, otherwise referred to as baseflow separation or hydrograph separation [40]. This method includes some manual estimation of springflow, making this method's results often subjective and time-consuming. The USGS often needs up to three months to publicly release springflow estimates, though the EAA and other water resource managers need to make daily water use decisions, especially when the aquifer levels are close to drought stage thresholds. Additionally, these springs are primarily modeled by the EAA using numerical models to simulate pumping and drought scenarios. Here, a springflow prediction model was created using a theory-guided data science approach as a viable option for water resource management.

Initialization of Model Parameters

Theory-guided initialization of model parameters was implemented when selecting predictor features for this thesis. Because there is no comprehensive dataset for all hydrological data in the Edwards Aquifer region, features were actively selected from federal and state agencies and included based on their proximity to Comal Springs and the hydrology of the Edwards Aquifer. Precipitation was gathered for the entire San Antonio region but only a subset of data was selected in the drainage area and recharge zone of the Edwards Aquifer because this is the source of precipitation that most likely impacts springflow. If this project was completed by a data scientist without background knowledge of the system, the predictor features

would likely have been selected differently. For example, without domain expertise in hydrology, a data scientist might select a subset of weather gauges within a 30 mile radius of Comal Springs and would not realize that precipitation that falls in the confined zone of the aquifer would not infiltrate and therefore not impact springflow discharge at Comal Springs.

In selecting parameters, the time frame for this study was taken into consideration. Data were included only if they were recorded with minimal breaks between the 1950 and 2019. Because the drought of record in the 1950s led to Comal Springs ceasing flowing has set precedent for many policy and law, it was essential to include this time in the study. This is the only time the springs were so low that measurements read as 0 $\frac{ft^3}{s}$. There are quite a few periods of time during which drought states were experienced (Figures 2.3 and 2.4) and these are the times when it is most critical to accurately predict springflow discharge.

Constrained Optimization and Feature Engineering

Theory-guided constrained optimization and feature engineering refers to the restriction of space of the model parameters [16]. Weather gauges or other monitoring systems can experience error and misrecord data that leads to impossibly large or small values. Using domain expertise, values from each feature were truncated to remove data that that are physically implausible, such as precipitation values less than zero inches per day.

Refinement of Data Science Outputs

Theory-guided refinement of data science outputs refers to the alteration of model outputs to make them more compliant with our understanding of the natural system [16]. From domain knowledge of the Comal Springs system, we know that

Comal river discharge is equal to Comal springflow discharge when there are no precipitation events. When precipitation occurs, river discharge will be larger than springflow due to accumulated runoff. Therefore, Comal springflow discharge can never be greater than Comal River discharge and this natural phenomena was set as a rule for post-processing of the model outputs. In each instance when springflow exceeded river discharge in the model outputs, springflow was truncated and set equal to river discharge. This refinement of modeled springflow outputs with the additional, corrective algorithm is key to producing springflow discharge estimates consistent with what is seen in nature.

Table 3.1: Problem Formulation for Springflow Discharge Estimation

Goal:	Predict springflow discharge from Comal Springs with high accuracy using theory-guided data science
Objective Function:	$min \sum_{i=0}^{n} w_i(y_i - \hat{y}_i)^2$
Subject to:	$Q_{spg} \leq Q_{riv}$
All Potential Predictor Features:	J-17, Q_{riv}, J-27, P_{RM}, T_{max}, ΔT, T_{min}
Response Feature:	Q_{spg}

Chapter 4

Machine Learning for Statistical Analysis of Comal Springs Discharge

The aspiration of this research to assure results are shareable and accessible for future work and, to achieve this, the principles and tenets that assure open source and reproducibility have been closely followed. The workflow followed in this research includes feature ranking/feature selection, isotonic regression, and an attempt to improve the model output with model boosting. Feature ranking and feature selection were included prior to model runs in order to understand the relationships between all the considered predictor features. This included summary statistics of the data and truncation of values that fall outside their natural range, univariate and bivariate analysis to visualize the data distributions and relationships, and model-based feature ranking with the least absolute shrinkage and selection operator (LASSO). Each metric alone may misrepresent the true feature importances. However, when

the results from all metrics are examined together, the data scientist can make an informed interpretation and feature ranking.

Isotonic regression was used as the machine learning model in this study to make predictions on springflow levels. This method involves fitting a curve by splitting it into k isotonic contraints, or segments, and fitting a linear model between each constraint. The isotonic regression model assumes that the slope is never decreasing. This model was chosen because of the change in linear slope at low levels of Comal Springs discharge and J-17 (Figure 4.13) more closely than that of a linear regression (Figure 4.12.

To improve upon the isotonic regression, model boosting was performed. A fast-learning, stacking approach was used through which the error from a strong model is calculated and then used to fit another model. This process is repeated until the desired accuracy is obtained. The isotonic regression is theory-guided in choice because of its ability to capture the change in linear slope in the model. The models used in an attempt to improve upon the isotonic regression were a multiple linear regression and naive Bayes classifier. The naive Bayes classification method is based on the conditional probability of a category, k, given n features, $x_1 \dots x_n$ and builds upon fundamental Bayesian statistics. However, model boosting with distinct features requires conditional dependence between features and this was not a correct assumption, as seen below when testing for conditional independence was performed. Therefore, model boosting was not an appropriate next step and the final results of this modeling workflow were obtained through the isotonic regression followed by theory-guided refinement of model outputs.

4.1 Introduction to Open Source Machine Learning Applications

Open source is defined by the Open Source Initiative [25] as a software or program that has (1) free distribution, (2) source code, (3) derived works, (4) integrity of the author's source code, (5) no discrimination against persons or groups, (6) no discrimination against fields of endeavor, (7) distribution of license, (8) license must not be specified to a project, (9) license must not restrict other software, and (10) license must be technology-neutral. The benefits of open source include fast bug fixes in the source code, a large online community, and broad adoption of the software or programming packages. The open source community has created reputable machine learning technologies, especially in Python. These include keras, scikit-learn, TensorFlow, and Theano. The Python library, scikit-learn, was used in this study as it contains prepackaged tools for data normalization and standardization, statistical analysis, machine learning model runs, and model output post-processing. All Python code written for this study is included in an online repository [26]. Every workflow in this study, from accessing data subsequent preprocessing to machine learning model runs, is completely reproducible and open source for other researchers to learn from and incorporate into their studies as desired.

4.2 Data Sources

When performing statistical learning methods on big data, there is often a precompiled dataset containing all parameters for analysis. Here, there was no master dataset of all Edwards Aquifer data and all data used in this study were individually accessed from their respective agencies and preprocessed prior to model runs

through theory-guided initialization of model parameters. The individual features selected to be included in the comprehensive Edwards Aquifer dataset are from the EAA, USGS, and National Oceanic and Atmospheric Administration (NOAA) (Table 4.2, Figure 4.2). The time frame for this study is from 1950-01-01 to present using daily data that captures the two major droughts in recent history (1950-1956, 2011-2014).

The EAA Uvalde (J-27) and Bexar (J-17) county index well data contain records of the daily maximum recorded well level in meters above sea level (msl). The J-17 index well is located in downtown San Antonio, TX (approximately 24 miles from Comal Springs) and the J-27 index well is located in downtown Uvalde, TX (approximately 106 miles from Comal Springs, Figure 4.2). These well levels are used by water management to determine aquifer health. The regional drought critical periods are set based on the water levels recorded at these two wells (Figure 2.2). Index wells J-17 and J-27 are triggers for the San Antonio pool and Uvalde pool of the Edwards Aquifer, respectively. Once the water level falls below a certain threshold over a 10-day average, the region experiences a critical drought period with groundwater withdrawal reductions enforced.

Daily streamgage data for Comal River (Q_{riv}) and Comal Springs (Q_{spg}) were downloaded from the USGS National Water Information Systems (NWIS) database (Figure 4.3) [37]. Comal River discharge (USGS Site 08169000 Comal River at New Braunfels, TX) was selected as a parameter in the model because of its proximity to the springs, located just downstream of the spring orifice. Comal River discharge data are calculated every 15 minutes from a rating curve using the most recent gauge height measurement then averaged daily by the USGS. Springflow discharge from Comal Springs (USGS Site 08168710 Comal Springs at New Braunfels, TX) is the

response variable being predicted in the study. The Comal Springs discharge data are back-calculated daily from the Comal River discharge data through hydrograph separation by the USGS. During periods of no precipitation, the Comal River gauge directly records springflow from Comal Springs, as this is the only upstream water source. During precipitation events, the USGS performs hydrograph separation to estimate both the volume of springflow and river discharge from runoff. After precipitation events, the volume of springflow is recalculated based on the Wahl and Wahl hydrograph separation method and updated [40].

The NOAA weather gauge data contain maximum temperature (T_{max}), minimum temperature (T_{min}), and precipitation (P) (Figure 4.5). NOAA has a national network of weather gauges [23], but the subset of gauges used here were selected based on their location in the aquifer region, i.e. intersection with the Edwards Aquifer drainage area and recharge zone (Figure 4.2). The NOAA daily weather summaries provided were recorded by 54 weather gauges with varying periods of record. Of the 54 weather gauges in this study, 50 are evenly distributed across the Edwards Aquifer drainage area and 4 are sparsely distributed across the eastern side of the recharge zone. Because there were so few weather gauges in the recharge zone, the data from all weather gauges were transformed and organized into a clean, tabular format, then averaged together to create one daily feature for T_{min}, T_{max}, and P.

4.3 Theory-Guided Feature Engineering

Two additional potential predictor features were created from the NOAA data: ΔT and rolling mean of precipitation. ΔT was calculated by subtracting the T_{max} from T_{min}. The rolling mean of precipitation (P_{RM}) was calculated with a trailing

window of $t = 3$ days. Precipitation has a value of 0.0 $in.$ in approximately $\frac{2}{3}$ of observations, which is expected in the region yet not useful when trying to make predictions. Therefore, P_{RM} replaced P as a predictor feature.

The predictor feature ΔT was created in order to determine how much of an impact the daily fluctuation in temperature has on springflow discharge, and whether it is more useful than T_{max} or T_{min}. The P_{RM} predictor feature was created because a large number of days per year in the region recorded $0in$ of precipitation. Taking the rolling mean of precipitation was done to smooth the signal, since it is common for the region to experience large periods of no precipitation followed by intense storm events, and it is not very useful to include a predictor feature that includes the same value ($0in$) for a large percentage of the dataset.

4.4 Feature Ranking

4.4.1 Model Initialization

Comal Springs discharge (Q_{spg}) is the response feature being predicted. This is done by estimating the function, $\hat{f}$, that describes springflow to solve for the following:

$$\hat{y} = \hat{f}(X_1, \ldots, X_n) \tag{4.1}$$

where:

$\hat{y}$ is the response feature

$\hat{f}$ is the estimated function

$X_1, \ldots, X_n$ are the predictor features.

Without eliminating any features through feature selection, the function that describes the springflow response is:

$$\mathrm{Q}_{spg} = \hat{f}(\mathrm{J17},\ \mathrm{J27},\ \mathrm{Q}_{RM}, \mathrm{P}_{RM}, \mathrm{T}_{max}, \Delta\mathrm{T}, \mathrm{T}_{min}) \tag{4.2}$$

4.4.2 Feature Selection for Isotonic Regression

Feature selection was performed in order to understand the relationships between the data available in the region. The metrics used here for ranking and data preparation are a combination of (1) summary statistics, (2) visual inspection of the data distributions and scatter plots, (3) correlation coefficients, (4) model-based feature ranking, and (5) expert knowledge. Data-driven metrics were combined with expert knowledge to ensure that the physics of the system are not ignored. After ranking the features and selecting which to include in this case study, the next step is to test and evaluate the predictive accuracy of the isotonic regression.

Summary Statistics

The summary statistics for this dataset can be found in Table 4.1. Values outside the typical physical range were truncated with the assumption that gauging stations malfunctioned and that outliers are not helpful in describing the system. Comal River was truncated at $\mathrm{Q}_{riv} < 800$ cfs and values less than 0 were removed from $\Delta\mathrm{T}$.

Visual Inspection - Univariate Analysis

Histograms for each feature are located in Figure 4.6. The P_{RM} has many small values which is expected in the dry, Texas climate. Index well J-17 follows a very

similar distribution to Q_{spg}. Index well J-27 is located further from Comal Springs than J-17 and exhibits a much different distribution, especially in low values.

The Pearson's product-moment correlation (r), Spearman rank correlation (r_s), and the partial correlation (pr) coefficients were calculated between each predictor feature and the Q_{spg} (Figure 4.10a). Each predictor feature's r and r_s have a relatively similar coefficient, indicating very few outliers in the predictor features. Index well J-27 has a positive r of 0.75 and a r_s of 0.79 but has a pr of -0.18, indicating that index well J-27 is negatively related to Comal Springs discharge values when the influences of confounding variables are removed (Figure 4.10a). P_{RM} has very small r and r_s values with a much larger pr of -0.35. Though the temperature features (T_{max}, T_{min}, and ΔT) all have low coefficients, T_{max} is more useful than the other temperature data according to the r and r_s (-0.20 and -0.21, respectively).

Visual Inspection - Bivariate Analysis

The Pearson's correlation scatter matrix (Figure 4.7) provides a visualization of the linearity in the bivariate relationships. The temperature predictor features all exhibit a low degree of linearity with springflow, with T_{max} showing the most linear relationship. The linear relationship between index well J-17 and Q_{spg} is the strongest among all the predictor features, but both Q_{riv} and J-27 also have a strong linear relationship with Q_{spg}. The remaining predictor features have little to no visible linear relationship with Q_{spg}.

Pearson correlation and Spearman rank correlation heat maps were used in addition to the correlation scatter matrix to visualize r and r_s (Figures 4.8 and 4.9). It is again clear through both heat map figures that index well J-17, Q_{riv}, and J-27 have strong linear relationships with springflow. The rank correlation coefficients

are often larger than the Pearson correlation coefficients because r_s is more robust to outliers than r.

LASSO Feature Ranking

The LASSO (least absolute shrinkage and selection operator), a regression analysis method, was used to perform feature ranking. The LASSO model contains two parts: the residual sum of squares and a shrinkage penalty (L1-Norm) (Equation 4.3). By tuning the λ hyperparameter, the shrinkage penalty forces the features to shrink to exactly zero in order of least to most useful as described through the LASSO model:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \underbrace{\sum_{i=1}^{n}\left(y_i - \left(\sum_{\alpha=1}^{m}\beta_\alpha x_\alpha + \beta_0\right)^2\right)}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{m}|\beta_\alpha|}_{\text{L1-Norm}} \tag{4.3}$$

where:

y_i is the observed springflow response

$\beta_{a...m}$ is the model coefficient

x_α is the predictor feature

λ is the LASSO shrinkage coefficient

LASSO ranked the standardized features in the following order from best to worst: (1) J-17, (2) Q_{riv}, (3) J-27, (4) T_{max}, (5) ΔT, (6) P_{RM}, and (7) T_{min}. Index well J-17 and Q_{riv} are much more informative than the other features (Figure 4.11a). To better visualize the feature ranking through LASSO, J-17 and Q_{riv} were removed and the process was repeated (Figure 4.11b). The T_{min} predictor feature shrinks to zero almost immediately.

Overall Feature Ranking

The overall ranking of features in predicting Comal Springs discharge is the following: (1) J-17, (2) Q_{riv}, (3) J-27, (4) P_{RM}, (5) T_{max}, (6) ΔT, and (7) T_{min}. Index well J-17 is the highest ranked feature in each of the metrics above. Index well J-27 and Q_{riv} ranked second or third depending on the metric, but Q_{riv} was interpreted to be ranked second overall due to its large r, r_s, and pr, with J-27 as third. Though the weather (P_{RM}, T_{max}, ΔT, and T_{min}) data performed low relative to the river and well data, T_{max} outperformed T_{min}, P_{RM}, and ΔT in LASSO and performed higher than the other weather data according to r and r_s. The data-driven approach ranked T_{max} higher than P_{RM} in some metrics, but through expert knowledge, P_{RM} has been ranked fourth because of its direct impact on aquifer recharge and Q_{riv}. T_{max} outperformed the remaining features and is ranked fifth. ΔT performed slightly better than T_{min} and, therefore, ΔT is sixth and T_{min} is least important.

4.5 Baseline Model: Linear Regression

Initially, a linear regression was attempted using index well J-17 as the predictor feature and Q_{spg} as the response feature. This baseline model performs well, with an R-squared value of 0.95. However, this model failed to capture the change in linear slope during drought conditions (Figure 4.12). The isotonic regression was next attempted to capture more of the variance in J-17 and springflow during periods of drought.

4.6 Isotonic Regression

Springflow discharge was predicted using J-17 as the single predictor feature through isotonic regression (Equation 4.4).

$$\hat{y}_i = min \sum_{i=0}^{n} w_i (y_i - \hat{y}_i)^2 \tag{4.4}$$

under the constraint that:

$y_0 \geq y_1 \geq \dots y_n$

where:

w_i are positive weights

y_i are real numbers

$\hat{y}_i$ are predictions

To test if the model was overfit to the data, the mean squared error (MSE) and variance explained (r^2) were calculated and visualized for a range of isotonic constraints (K) using k-Fold cross validation with 5 folds. However, any K value between 10 and 30 results in approximately the same amount of variance explained ($0.960 < r^2 < 0.961$). It is important for water management to accurately predict springflow values during drought conditions when discharge values are low. Therefore, in order to avoid overfit and capture the change in linear slope at small values of J-17, an isotonic constraint of $K = 15$ was selected (Figure 4.13).

The isotonic regression produced an r^2 of 0.96, meaning that 96% of the variance in springflow can be explained by index well J-17 (Figure 4.13). When comparing the measured springflow to the modeled springflow ($y - \hat{y}_{J17}$), the difference in the mean was 0.52 and the standard deviation was 18.43 with a minimum

value of -65 and a maximum value of 84. When measured and modeled data are plotted, they result is highly linear with a slope of 0.96 (Figure 4.14). When viewing the measured and modeled data in a time series, the model nicely captures the overall trend, but struggles in some cases to predict at the peaks (Figure 4.16a). The residuals from the model can be seen in Figure 4.15, and the modeled vs. measured springflow in a time series can be seen in Figure 4.16a.

4.7 Conditional Independence Between Springflow Residual and Predictor Features

To assess how well the residuals from the isotonic regression could be explained by the remaining predictor features, the conditional probability distributions between each predictor feature (J-27, Q_{riv}, T_{max}, T_{min}, ΔT, and P_{RM}) and the isotonic residual ($y - \hat{y}_{J17}$) were examined. If conditional dependence exists between the residual and the remaining predictor features, model boosting with additivity of components could be used to improve the predictive accuracy of the model. The marginal and joint probability distribution functions provide evidence that there exists conditional independence between features, meaning that the features do not explain the signal from the isotonic regression residuals (Figure 4.17). According to the mutual information or a measure of dependence between variables, index well J-27 and Q_{riv} were the two most informative features in understanding the residuals (Figure 4.18). The remaining predictor features (T_{max}, T_{min}, ΔT, and P_{RM}) each had a mutual information score below 0.07 and were not important in explaining the springflow residual. The conditional statistics (P10, P90, and expected value) for each predictor feature were visualized to determine whether the predictor features were conditional independent to the isotonic residual (Figure 4.19). The general

trend of most of these features' conditional statistics is constant, meaning that as the predictor feature increases, there is no change in expected residual value.

Violin plots were generated to visualize the conditional statistics with the left side (low) of each violin being less than the P50 value of the residual and the right side (high) being greater than or equal to the P50 value (Figure 4.20). The P50 value of the high and low residuals are all very similar. The P25, P50, and P75 values for both the low and high residuals are also very similar, further indicating conditional independence.

From the above metrics, it is clear that the predictor features are conditionally independent from the isotonic residual, and further modeling through boosting, an ensemble of weak prediction models, will likely not result in improved predictive accuracy. To determine if this was the case, additional models were run on the isotonic residuals. A multiple linear regression and naive Bayes classifier were used to create a boosting model in an attempt to improve upon the isotonic regression. The boosting method did not succeed in improving upon the isotonic regression, as was suggested by the conditional independence metrics. With an r^2 of 0.96, it can be argued that the isotonic regression is a strong model and there is limited remaining signal to explain.

4.8 Isotonic Regression with Uncertainty Analysis

The isotonic regression was used as the final prediction model, predicting Q_{spg} from index well J-17 with $K = 15$ isotonic constraints and theory-guided refinement of outputs using Q_{riv}. Theory-guided refinement of outputs involves post-processing the model output to align results with our understanding of the phenomena.

To calculate the uncertainty of the model outputs, the data was binned by

isotonic constraint and summary statistics were calculated for each bin [26]. The box plot presents the range in predicted values in each isotonic constraint 'bin' (Figure 4.21) and the EAA critical periods are displayed, as provided in Figure 2.2. There is a 50% chance of the residual from the isotonic regression falling inside the boundaries of the boxes. The bottom and top of the boxes represent the P25 and P75 values. The interquartile range (IQR = 1.5 x (P25 + P75)) is represented by the whiskers, or tips extending from the boxes. The range in predictions is very small for the low values of isotonic constraints, which displays that the model performs well during drought conditions. The isotonic constraint 'bin' that represents the largest values of index well J-17 has the largest error range, indicating that the physics of the system are not as well captured through the isotonic regression during extremely wet periods (Figure 4.21). These events are much more rare and therefore difficult to predict. There are other sources of uncertainty. The accuracy of the springflow values used in the training data are assumed to be correct. Any biases or errors in this data will reflect in the model predictions. The use of a bootstrap method could also be used as an alternative in order to calculate multiple complete models for overall model uncertainty. This could be explored in future work.

The predicted springflow values were then corrected using theory-guided refinement of outputs to meet the condition that $Q_{spg} \leq Q_{riv}$. In Comal Springs, Q_{spg} is never greater than Q_{riv} because the springs directly feed the river, though river discharge is often larger than springflow during precipitation events. If the predicted springflow value was larger than the measured daily mean river discharge, springflow was truncated and set equal to river discharge. Additionally, if the Q_{spg} prediction was less than 0 cfs, the prediction was automatically increased and set equal to 0 cfs. The accuracy of the model was calculated by determining whether

an observation falls within the P25 to P75 range (between the bottom and top of each box in the boxplot). If the observation fell within that range, then it was correct. If it fell outside the range, then it was incorrect. This calculation was done prior to and after truncating springflow to fall within the physical range (Figures 4.21 and 4.22). The predictive accuracy of the model was calculated for each of the drought stages as well, using index well J-17 as the proxy. The results prior to and after correction are in Table 4.4. The predictive accuracy increases as the drought condition worsens for both methods, but there exists a clear increase in predictive accuracy when the springflow value is corrected.

4.9 Results for Predicting Springflow Discharge

The ability to make predictions with the machine learning model using real-time data is what makes this model useful to water management in estimating springfow values.The model performs well when making predictions solely from the value of J-17, and Q_{riv} is required for refining results with the theory-guided methods. Real-time, daily values for both J-17 and Q_{riv} are openly available and can be used to generate a springflow discharge prediction. Therefore, the approach requires minimal inputs (J-17 level and Comal River discharge) to create real-time predictions which can be automatically downloaded from each agency's website [5][9]. A codebase with steps to automate the data download and ingestion steps is available in a Github repository setup to share code used in analyses to complete this thesis [26]. Through the analytical workflows presented in Figure 4.1, data is gathered and formatted to be used as the input for the predictor input data. The real-time daily observation is then run through the isotonic regression and its result is subsequently corrected to not exceed the daily Q_{riv}. The model can be set up to automatically

retrieve the most recent daily data from each agency website, reformat it and input it into the model. However, a user could also use hypothetical values to generate a springflow prediction. Additionally, the code base is portable across case study sites and could be applied to another spring location.

4.10 Discussion

Here, the goal was to create a predictive machine learning model to estimate springflow discharge from Comal Springs using hydrological and meterological data in the region. Feature ranking was performed and through this process, it was discovered that index well J-17 (located 26 miles from the spring orifice) is highly correlated with Comal Springs discharge. To avoid multicollinearity and increase model stability, J-17 was used as the single predictor feature to model springflow using isotonic regression. Model boosting using multiple linear regression and naive Bayes classification was attempted with the isotonic regression residuals in an attempt to explain more of the variance, but was unsuccessful due to conditional independence between the residual and remaining predictor features.

Next, the model accuracy was examined and characterized for each of the drought stages set by the Edwards Aquifer Authority in the San Antonio region. Theory-guided post-processing was performed to create predictions that are consistent with the physical processes. For example, the relationship between springflow and river discharge provides a clear and consistent rule and constraint for any analysis (e.g. springflow can never be larger than total river discharge). Therefore, if springflow exceeded the daily river discharge, it was truncated and set equal to the discharge value. After this correction, the accuracy was calculated (Table 4.4), with a "correct" springflow value if the prediction fell between the P25 and P75 value (see

Figure 4.21). The ability to predict springflow with increasing accuracy as drought worsens is of benefit to water resource management. San Antonio groundwater withdrawal limits are put in place depending on the well levels at J-17, J-27, and Comal Springs. Making accurate predictions when Texans face drought conditions is more beneficial than predicting springflow when the springs are plentiful.

Domain expertise of the physical system is critical in order to produce viable springflow predictions. This springflow prediction model performed well prior to the correction and extremely well after the correction. There is high predictive accuracy when springflow is corrected, with a high chance of the predicted value falling within $\pm 10 \frac{ft^3}{s}$ of the actual springflow value (Figure 4.21). When the region experiences drought conditions and the recorded level at index well J-17 is within the drought stages set by the EAA, the model is highly accurate, which is when this model is most useful to water management. It is less essential to know how much water is coming from the springs when they are heavily flowing.

In this thesis, model boosting with multiple linear regression and naive Bayes classification was attempted to improve upon the isotonic regression. This was unsuccessful due to the residual from the isotonic model being strictly unstructured noise. The LASSO and multiple linear regression were attempted prior to the isotonic regression to assess their applicability in this specific problem but did not perform as well as the isotonic regression.

These are novel methods that require further examination prior to implementation or replacement of current springflow measurement methods. Due to the amount of averaging done by the USGS in their prediction workflow, this machine learning alternative is arguably as accurate as their current method. The Comal River discharge is measured every 15 minutes through a rating curve, but is averaged

daily and then used by the USGS to calculate springflow discharge. Additionally, their method cites the use of a computer program that follows the Wahl and Wahl method [40], but in reality they include some manual estimations. The method presented here requires no manual aspect to the workflow, making it faster and cheaper than the current method. However, it does mimic the data and predictions that have been made in the past in order to make new springflow predictions, meaning that any biases in the springflow predictions in the past are incorporated into this model.

Table 4.1: The statistical summary of all predictor and response features.

	Mean	Std	Min	25%	50%	75%	Max
J-17	663.11	17.41	612.51	650.95	664.7	676.35	703.31
J-27	866.98	16.24	810.95	863.7	872.35	877.43	889.05
Q_{riv}	281.18	99.76	5.5	218.0	290.0	347.0	793.0
P_{RM}	0.09	0.2	0.0	0.0	0.0	0.08	4.18
T_{max}	78.11	14.28	6.0	69.0	80.0	90.0	109.0
T_{min}	53.49	15.22	-1.0	41.0	56.0	67.0	84.0
ΔT	24.62	8.87	-40.0	19.0	24.0	31.0	69.0
Q_{spg}	276.23	94.61	0.0	215.0	290.0	345.0	513.0

Table 4.2: Summary and origin of Edwards Aquifer Data used in this study.

Model Feature	Data	Units	Source
Predictors	Comal River Discharge	$\frac{ft^3}{s}$	USGS
	Bexar County Index Well (J-17)	msl	EAA
	Uvalde County Index Well (J-27)	msl	EAA
	Precipitation Daily Summary	in	NOAA
	Maximum Temperature Daily Summary	$^\circ F$	NOAA
	Minimum Temperature Daily Summary	$^\circ F$	NOAA
Response	Comal Springs Discharge	$\frac{ft^3}{s}$	USGS

Table 4.3: Summary statistics for the residual between measured springflow and modeled springflow in isotonic regression

	Before correction	After correction
Mean	1.0	7.0
Std	18.5	13.1
Min	-69	-56
P25	-12	0
P50	-1	0
P75	12	12
Max	83	83

Table 4.4: Springflow prediction accuracy before and after correction. A prediction is considered correct if it falls within the interquartile range. The correction was calculated using theory-guided refinement of data science outputs.

	# Data	Accuracy (Before Correction)	Accuracy (After Correction)
Drought Stage 1	4326	52.9	81.6
Drought Stage 2	2517	52.4	82.5
Drought Stage 3	1089	55.1	86.6
Drought Stage 4	384	59.1	88.5
Drought Stage 5	109	76.1	95.4

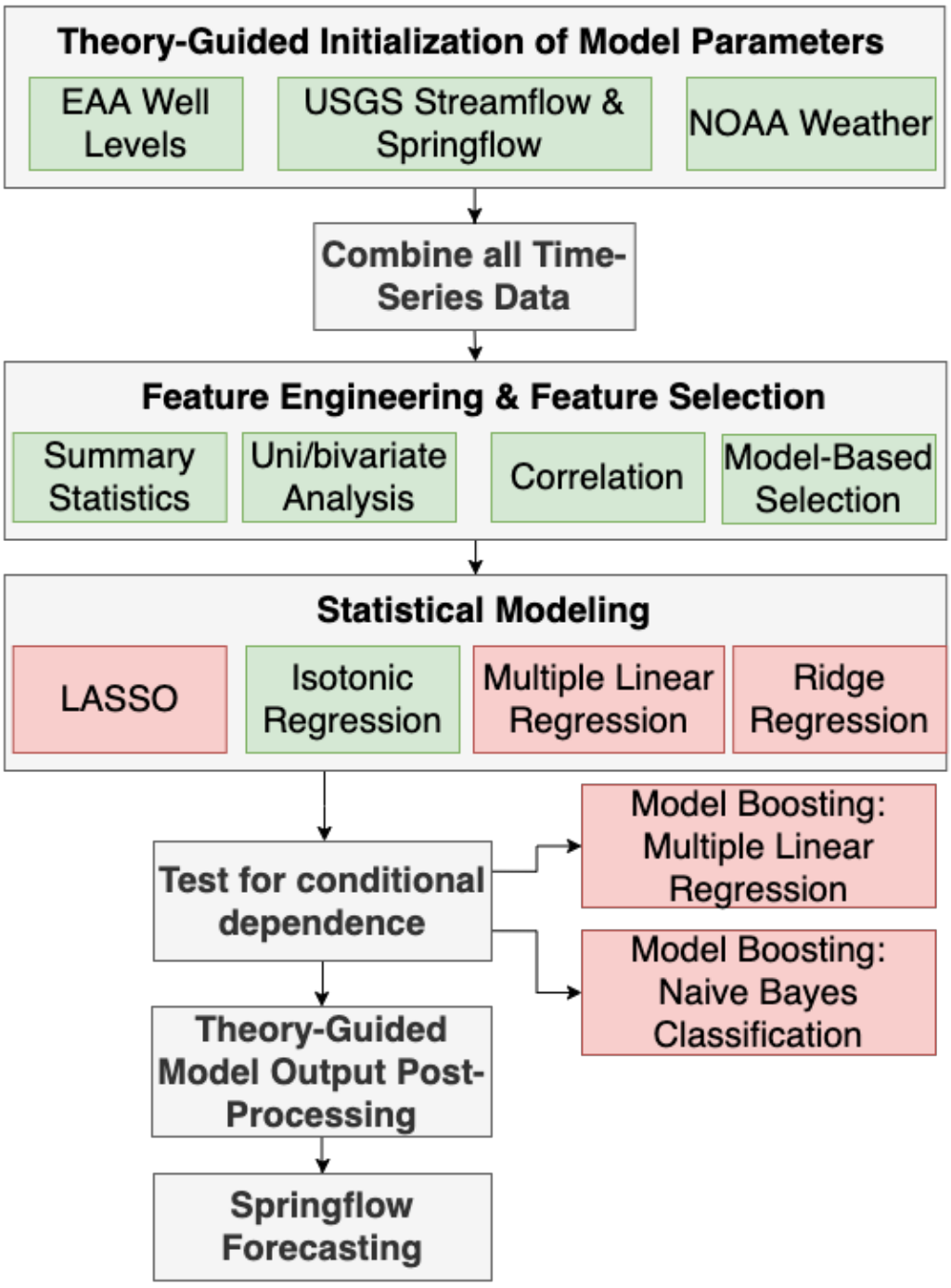


Figure 4.1: The full workflow used in building this springflow predictive model. The green boxes indicate the final path while red indicates explored paths. This demonstrates that various models were explored, but using the isotonic regression was the best solution for this predictive model.

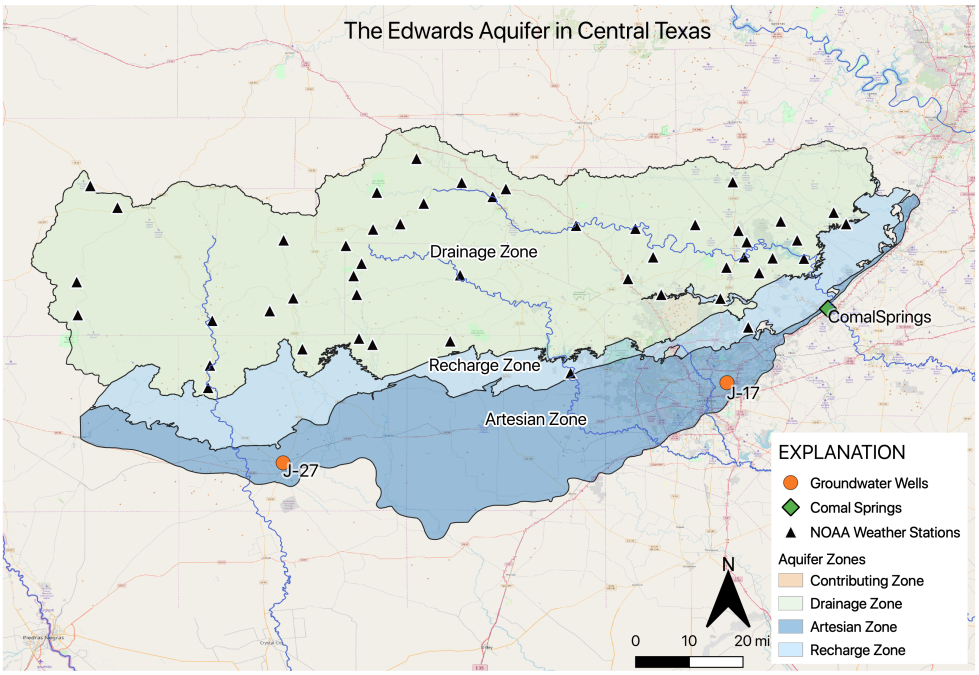


Figure 4.2: The Edwards Aquifer region with the locations of J-17 index well, J-27 index well, NOAA weather gauges, and Comal Springs indicated.

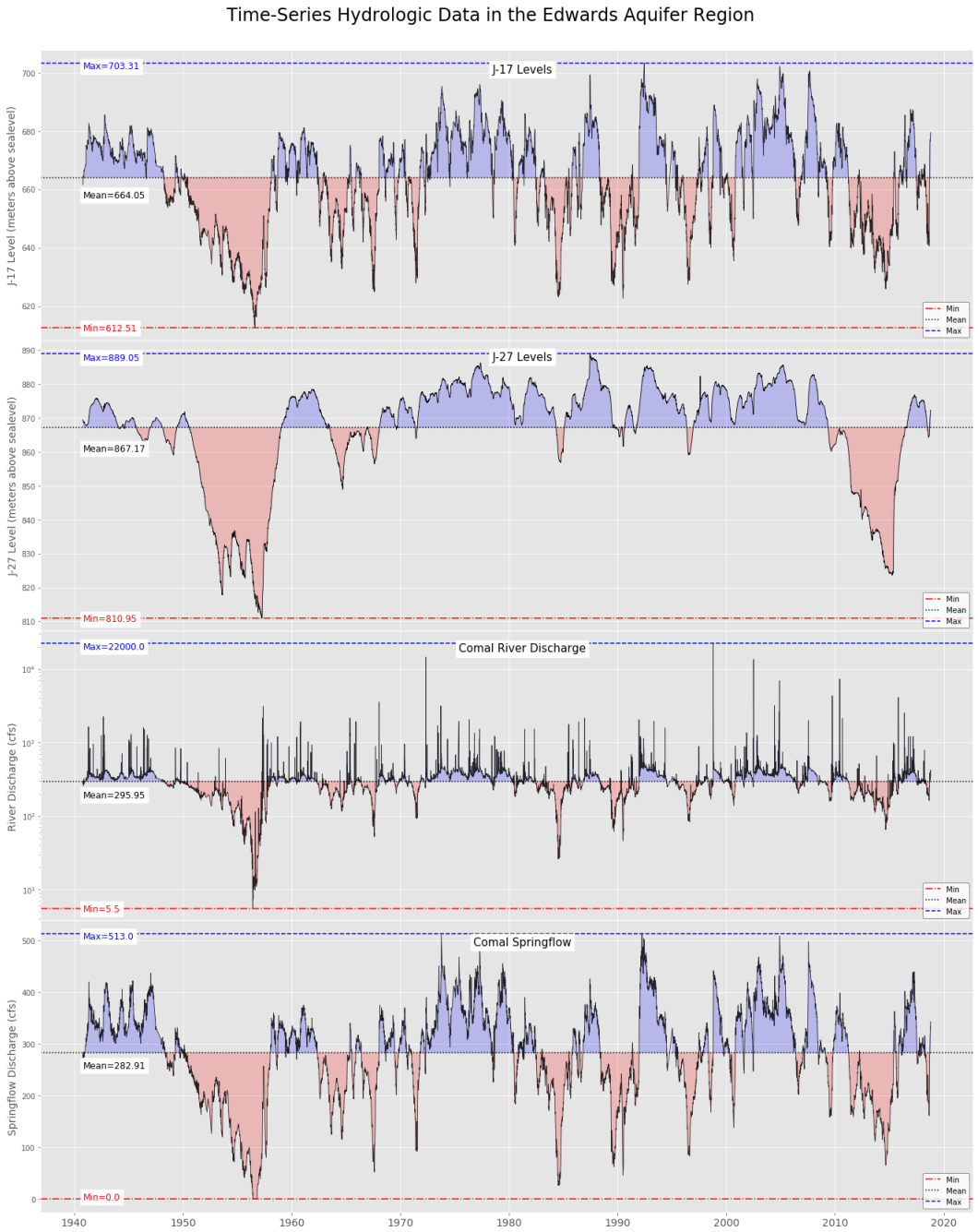


Figure 4.3: Hydrologic data of the Edwards Aquifer from 1940-2019. The time series data for Comal Springs, Comal River, J-17, and J-27 are plotted, with blue representing data above the mean and red representing data below the mean for each category.

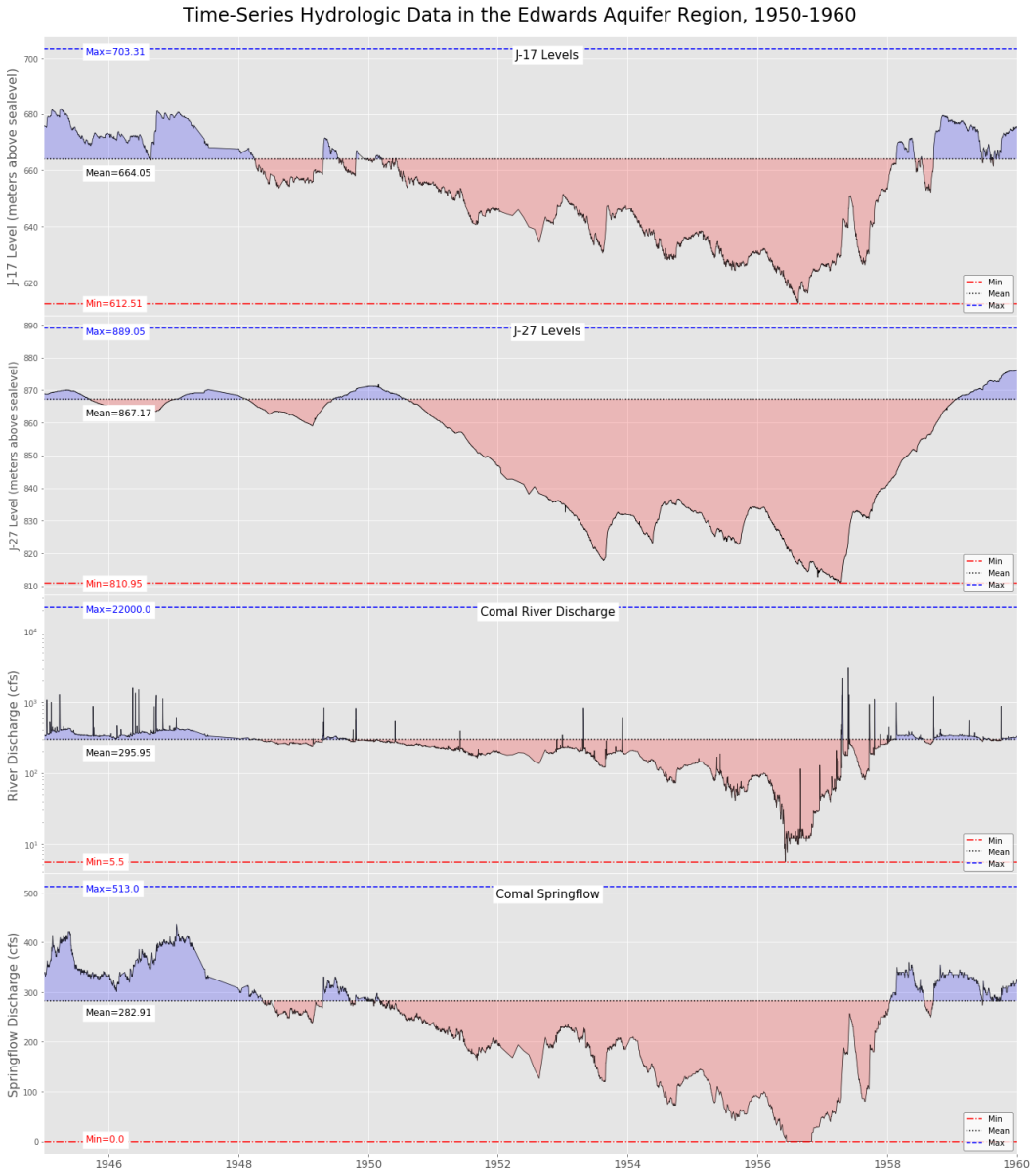


Figure 4.4: Hydrologic time-series data for Comal Springs, Comal River, J-17, and J-27 during the 1950s drought of record, with blue representing data above the mean and red representing data below the mean for each category. This is the only period of time during which Comal Springs discharge was recorded as 0 cfs.

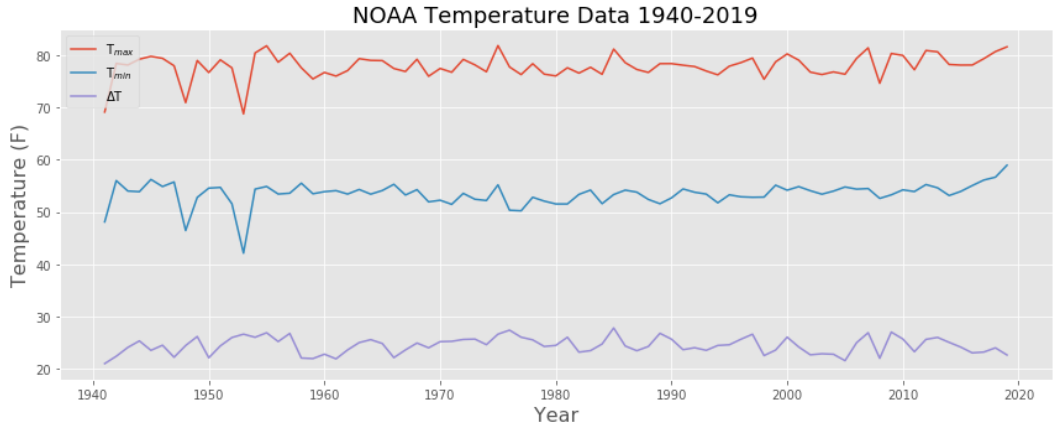


Figure 4.5: NOAA weather data (T_{max}, T_{min}, and ΔT) in a time series from 1950 to 2019.

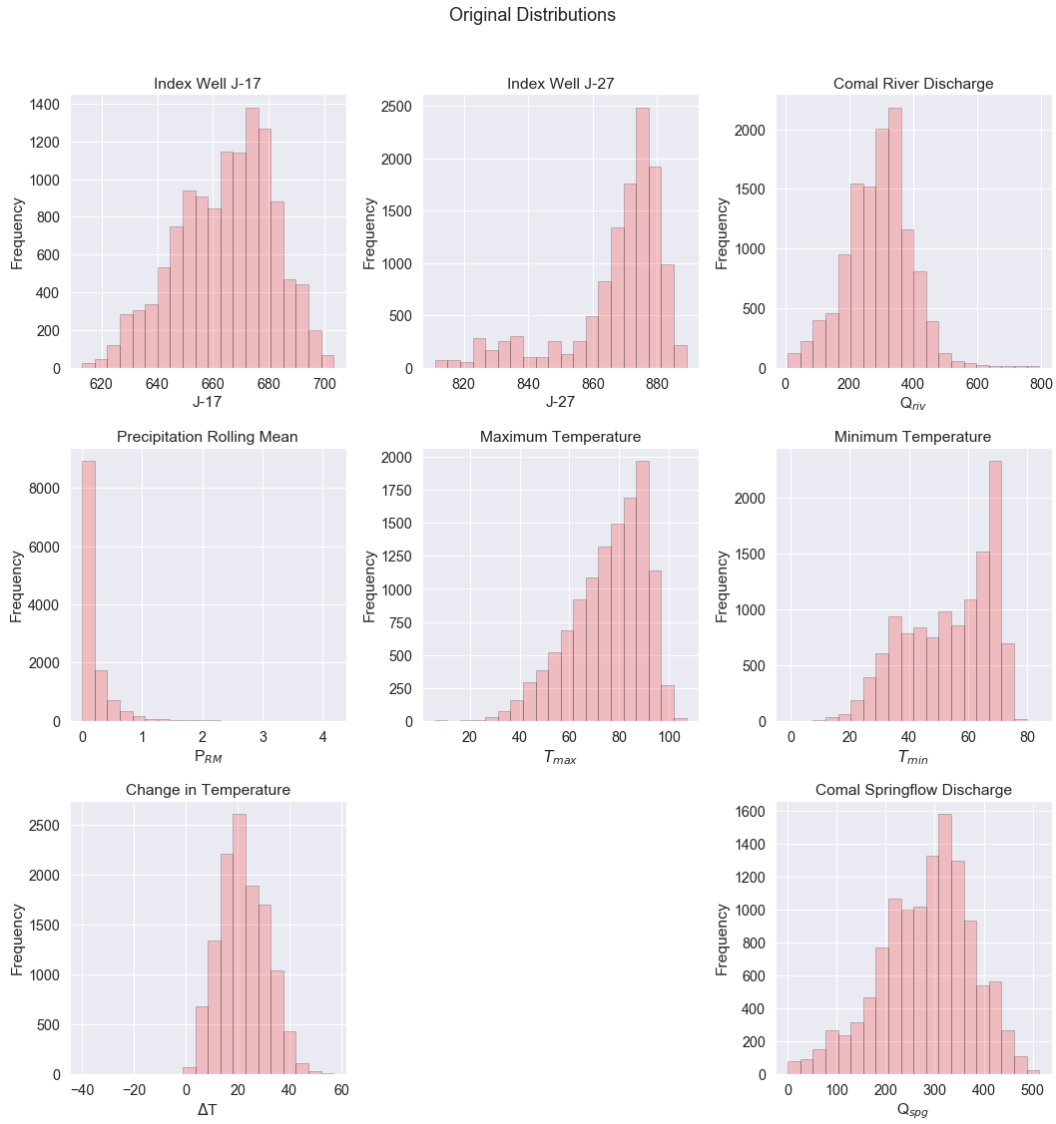


Figure 4.6: Histograms of the univariate, original distributions in each of the predictor and response features.

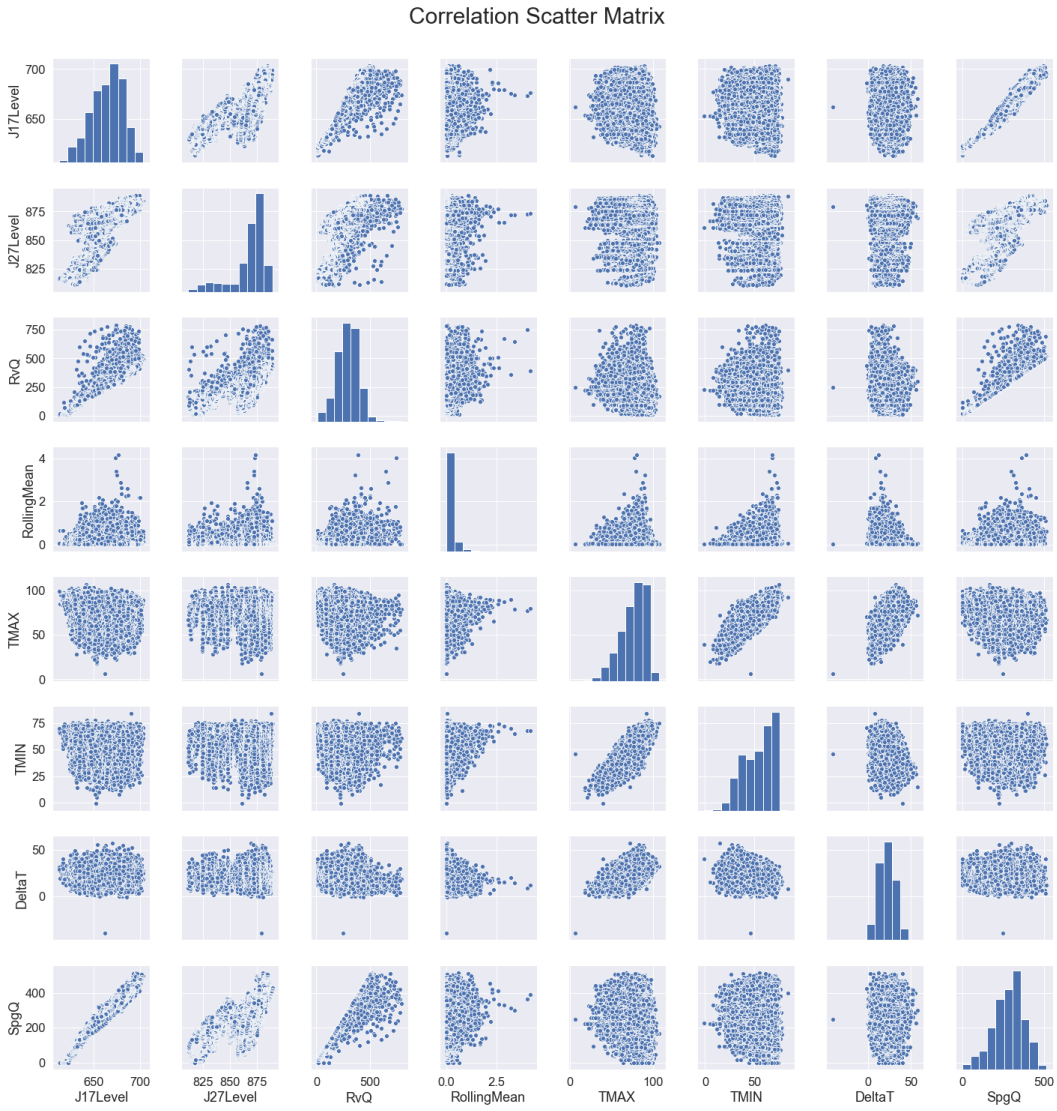


Figure 4.7: Pairwise correlation scatter matrix of the original data.

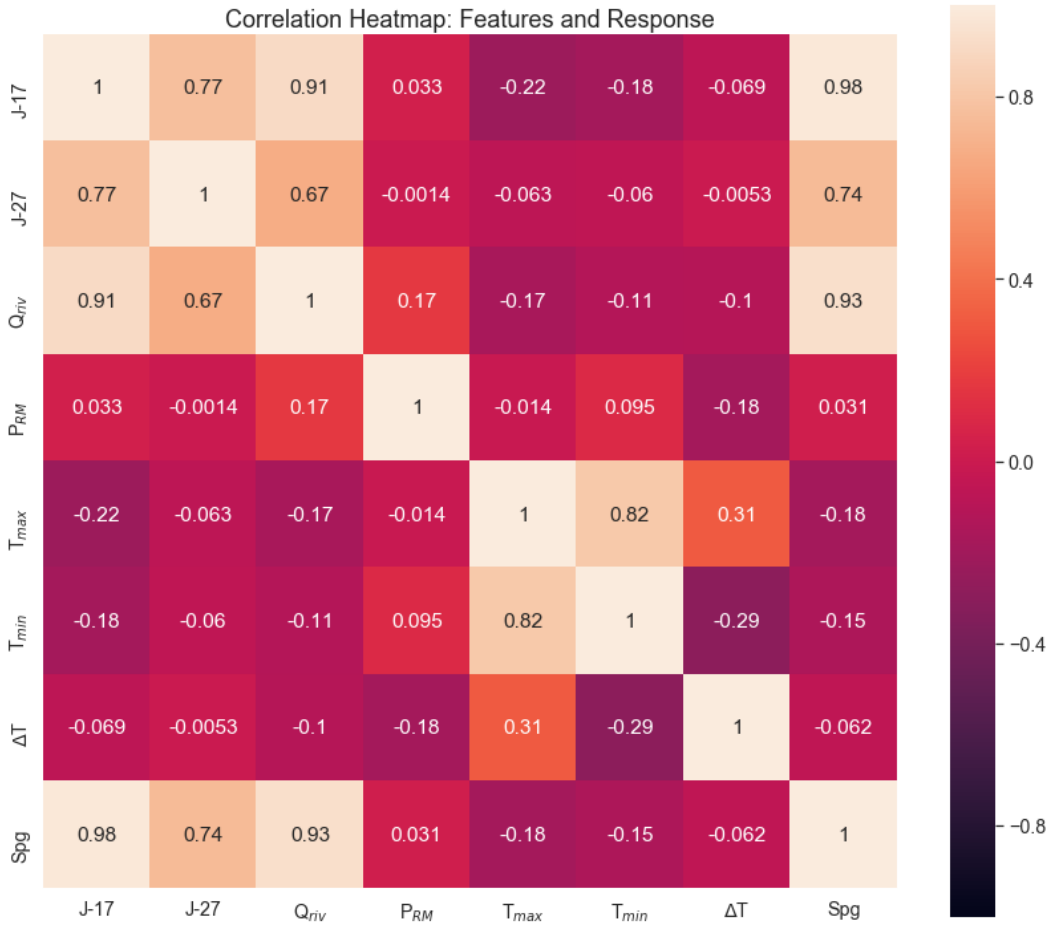


Figure 4.8: Correlation pairwise heatmap of the predictor and response features to visualize the degree of the Pearson product moment correlation coefficients.

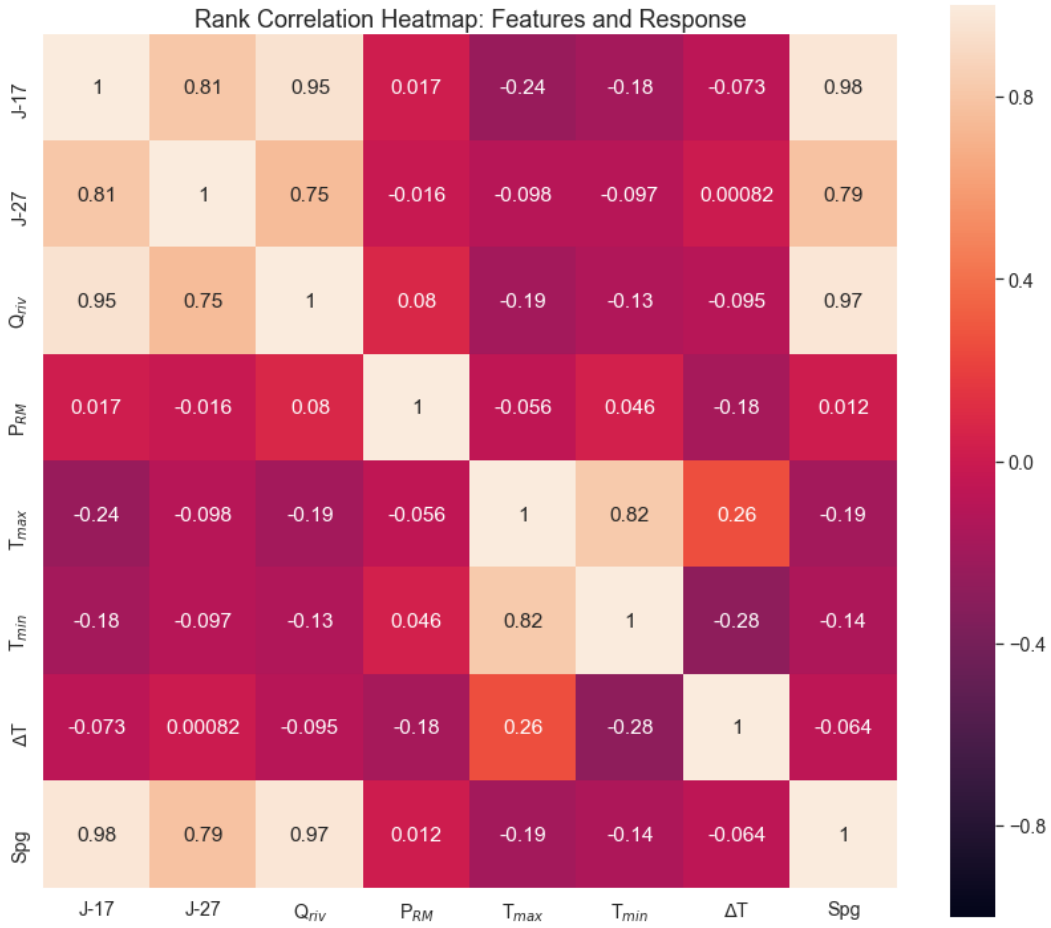
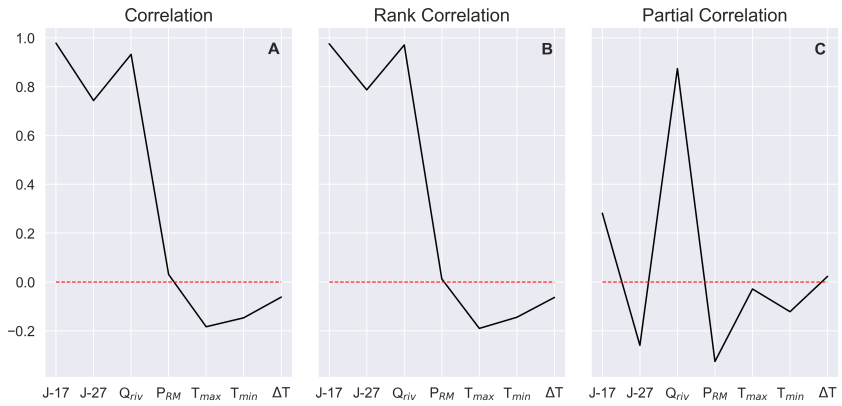
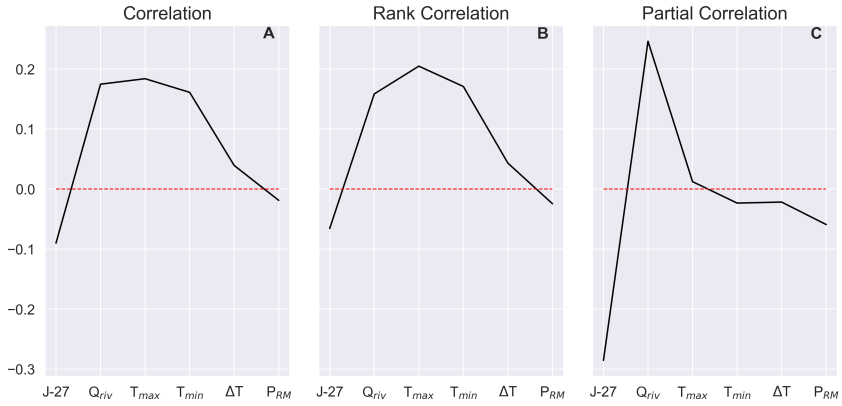


Figure 4.9: Rank correlation pairwise heatmap of the predictor and response features to visualize the degree of the Spearman rank correlation coefficients.

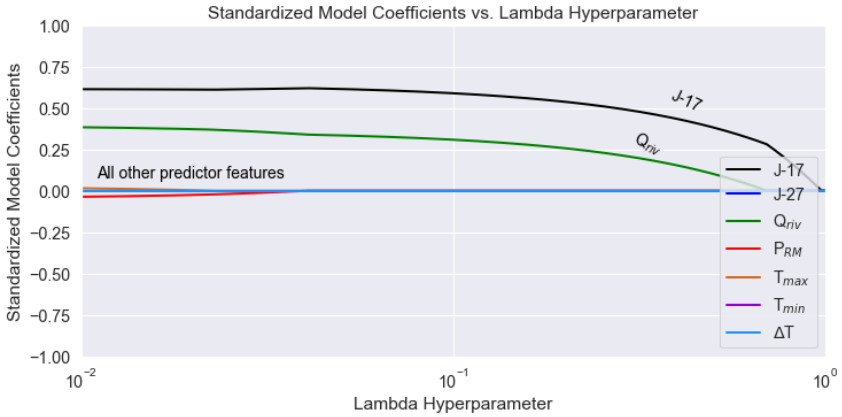


(a) Correlation coefficients for first boosting model: isotonic regression

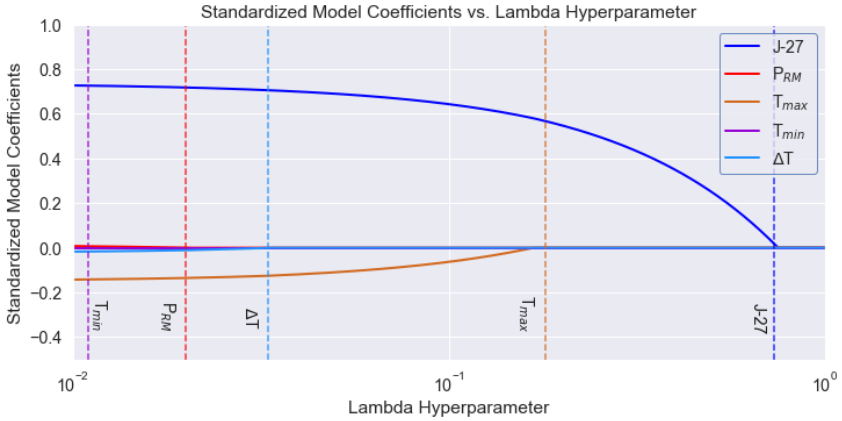


(b) Correlation coefficients for second boosting models

Figure 4.10: (a) Feature selection for isotonic regression with Q_{spg} as the predictor feature. (b) (a) Feature selection for the second boosting model (multiple linear regression, naive Bayes classification) with $y - \hat{y}_{J17}$ as the predictor feature



(a) LASSO model-based feature selection for isotonic regression with all predictor features.



(b) LASSO model-based feature selection for isotonic regression with index well J-17 removed.

Figure 4.11: (a) All predictor features ranked in order of importance for predicting Q_{spg}. (b) Index well J-17 was removed to better visualize and interpret the model-based feature ranking for predicting Q_{spg}.

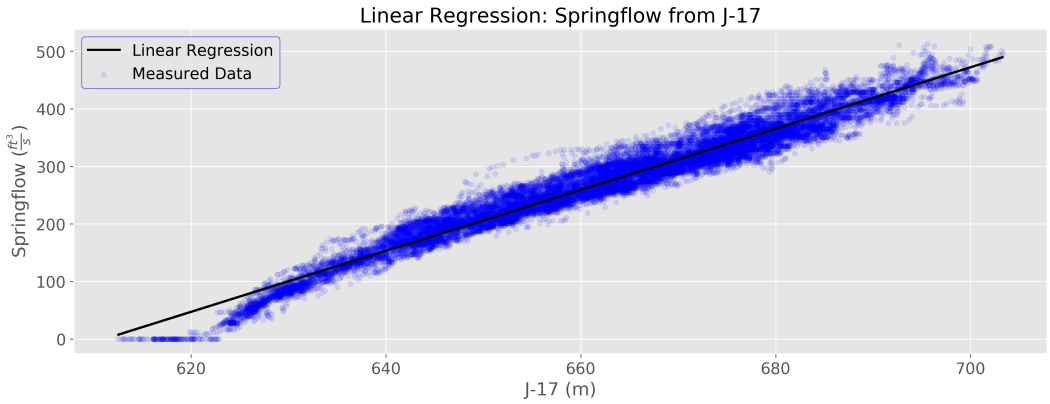


Figure 4.12: Linear regression using index well J-17 to predict springflow. This model performs well at high levels of J-17, but fails to capture the change in linear slope during lower well levels.

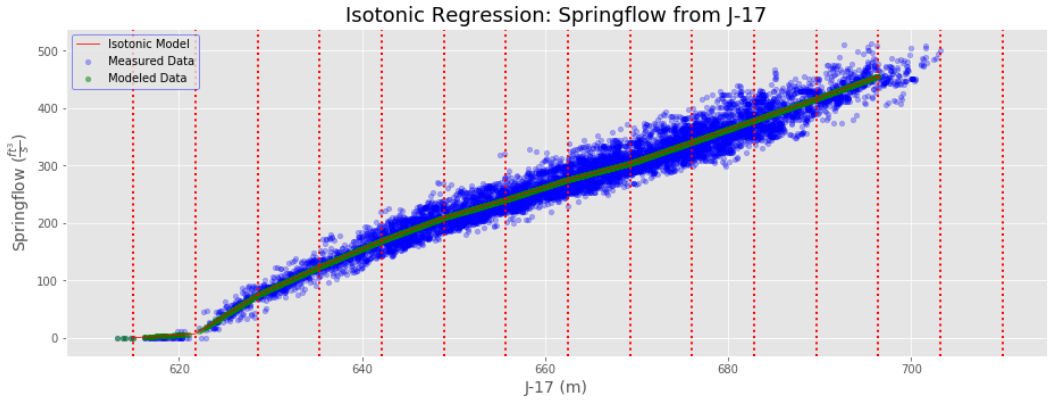


Figure 4.13: Isotonic regression model to predict springflow discharge from index well J-17. Variance explained from this model is 0.957.

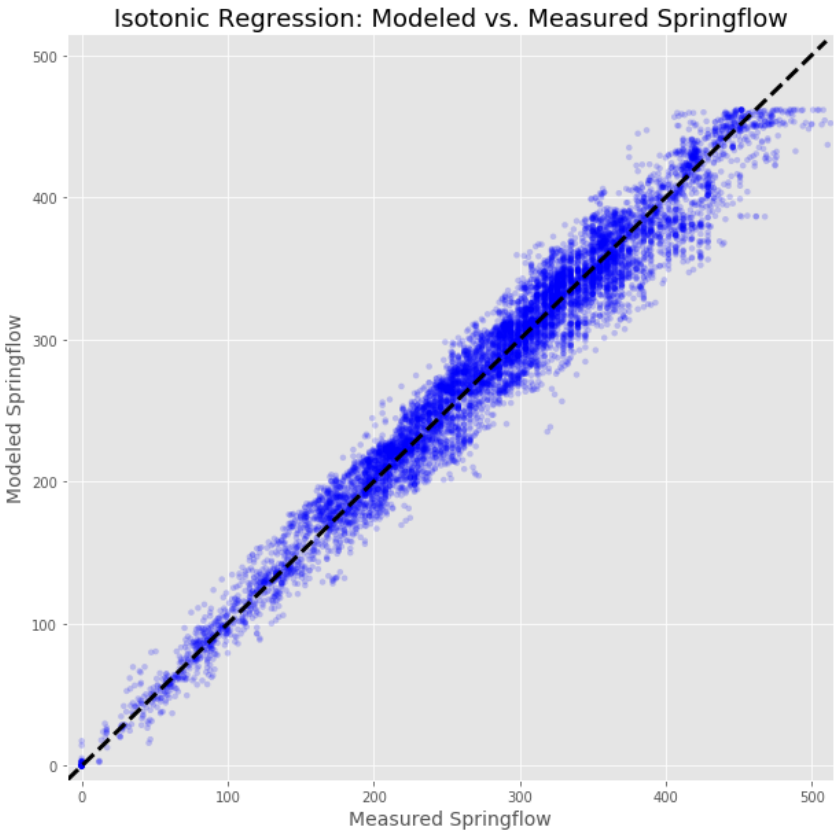


Figure 4.14: Comparison between the measured and modeled springflow through isotonic regression.

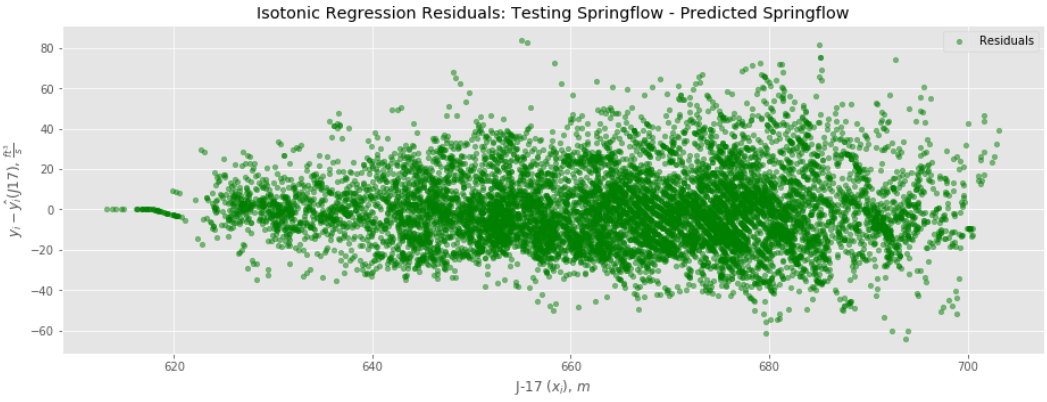
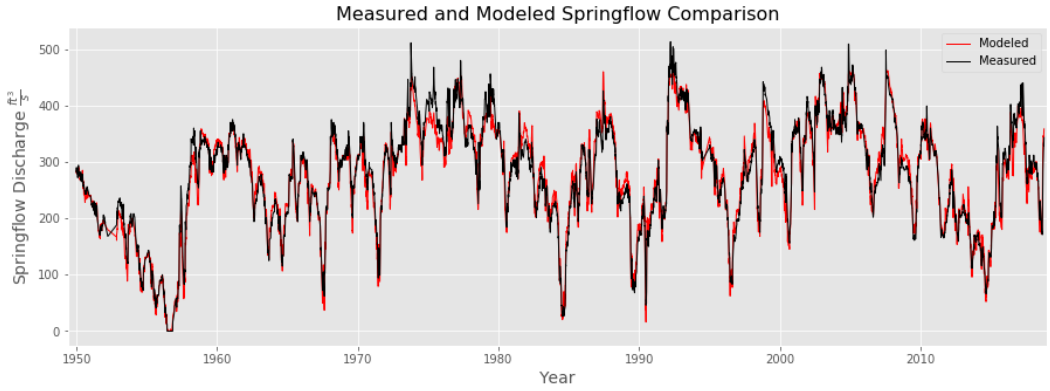
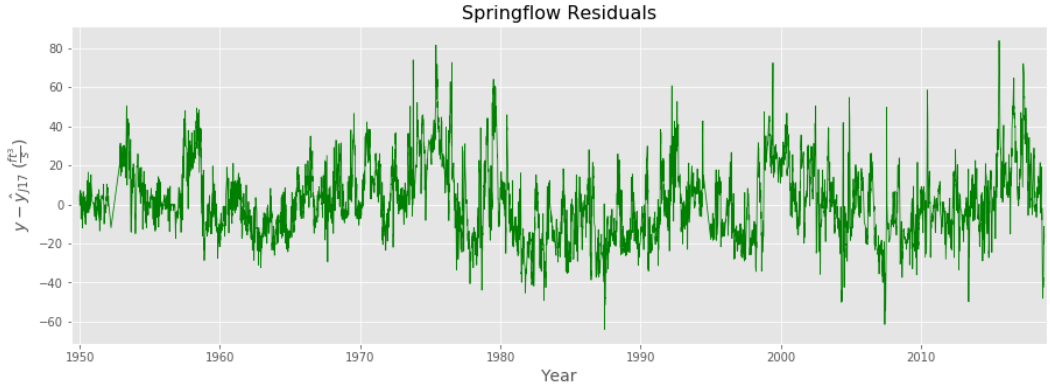


Figure 4.15: Isotonic regression residuals $(y_i - y(\hat{x}_i))$ using index well J-17 as the single predictor feature.

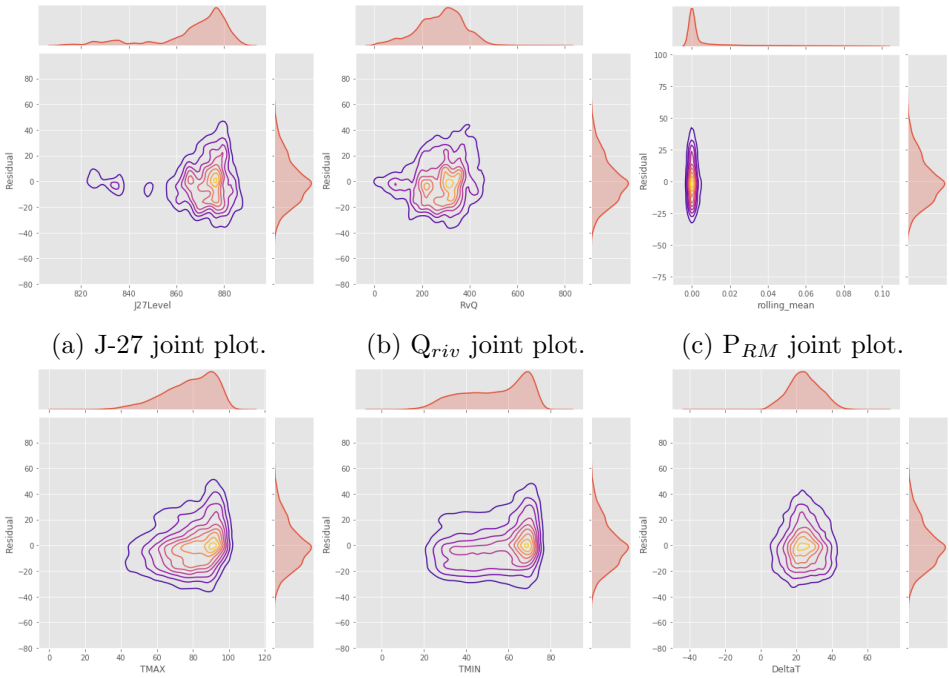


(a) The measured and modeled Q_{spg} on a time series.



(b) The Q_{spg} residuals ($y - \hat{y}_{J17}$) on a time series

Figure 4.16: (a) The measured and modeled Q_{spg} from 1950 to present, where red represents the modeled Q_{spg} and black represents the measured Q_{spg}. (b) The Q_{spg} residuals on a time series. The model struggles to accurately predict the peak events, but successfully follows the overall trend.



(a) J-27 joint plot. (b) Q_{riv} joint plot. (c) P_{RM} joint plot.

(d) T_{max} joint plot. (e) T_{min} joint plot. (f) ΔT joint plot.

Figure 4.17: The marginal and joint probability distribution functions for each predictor feature are displayed to test whether each predictor feature was conditionally independent with the isotonic residual.

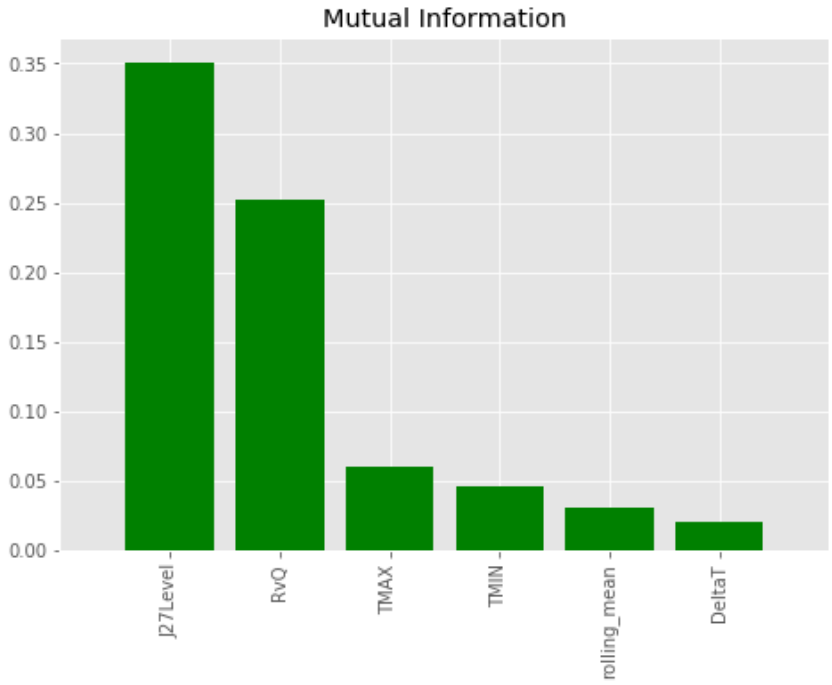
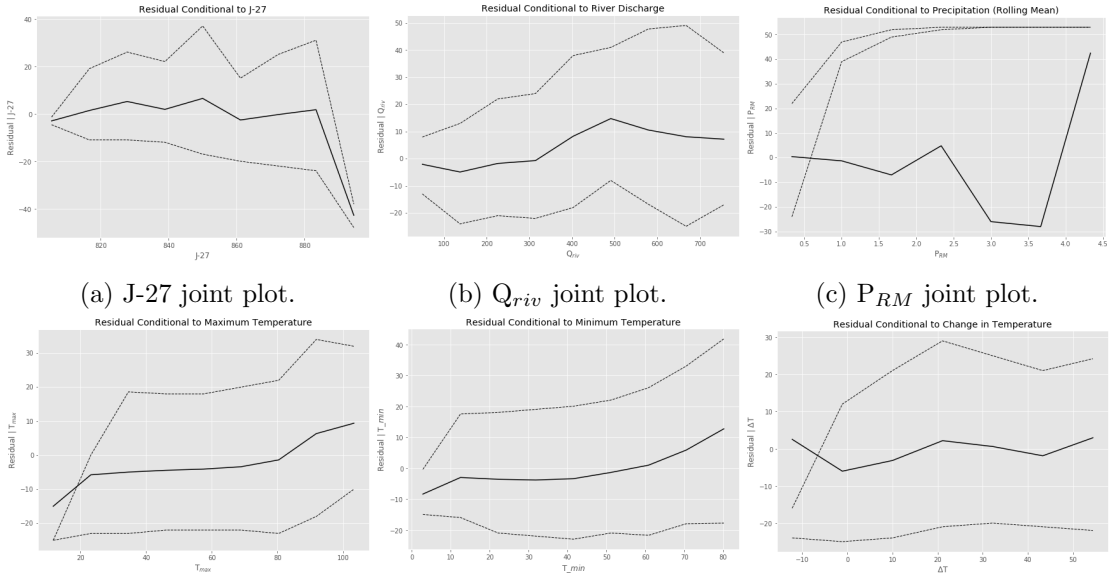


Figure 4.18: Mutual information of the isotonic regression residual and each remaining predictor feature. Index well J-27 and river discharge share the most information, but the weather data shares very little mutual information with the residual.



(a) J-27 joint plot. (b) Q_{riv} joint plot. (c) P_{RM} joint plot.

(d) T_{max} joint plot. (e) T_{min} joint plot. (f) ΔT joint plot.

Figure 4.19: The marginal and joint probability distribution functions for each predictor feature are displayed to test whether each predictor feature was conditionally independent with the isotonic residual.

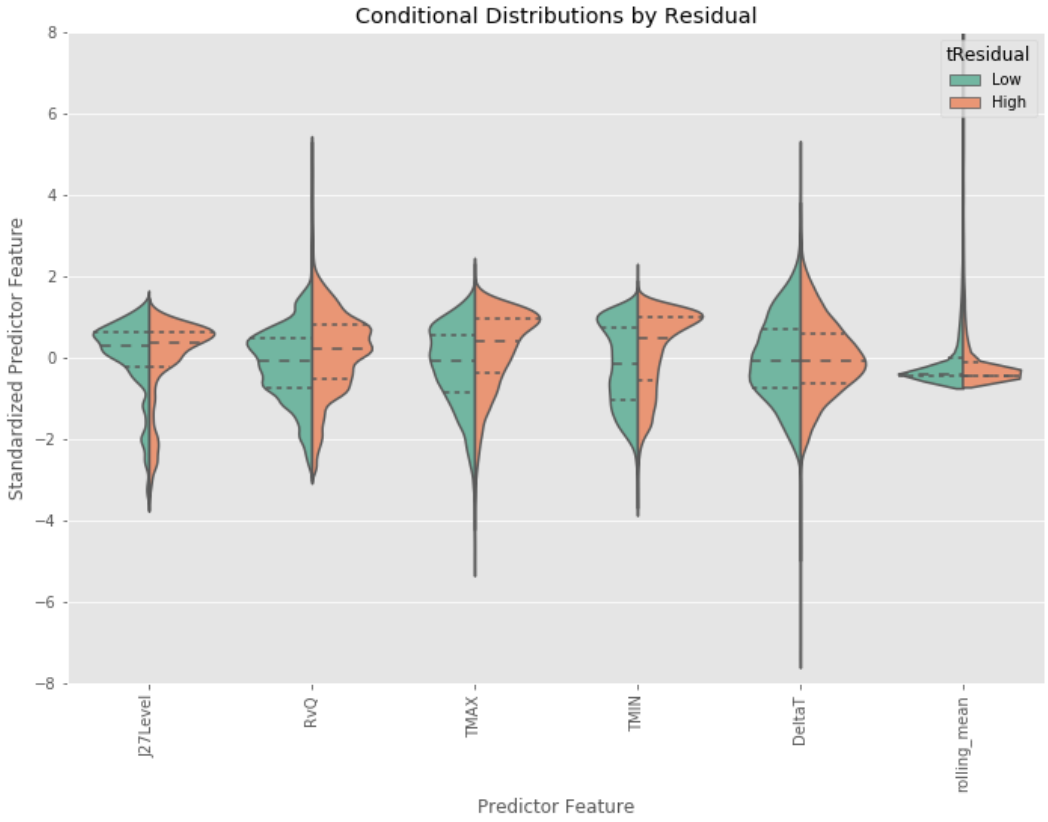


Figure 4.20: Violin plot to visualize the conditional distributions between the residual and each predictor feature. Most of the dashed lines (P25, P50, P75) are relatively equal and near zero, indicating conditional independence with the isotonic residual.

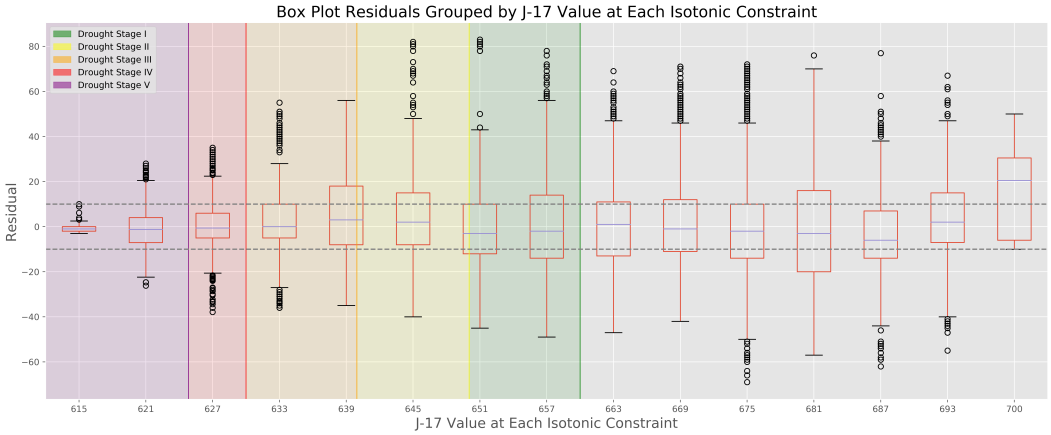


Figure 4.21: Box plots of the residuals plotted against the J-17 value at each isotonic constraint to display the distribution of predictions and the range of error. The colors represent the EAA critical periods (see Figure 2.2).

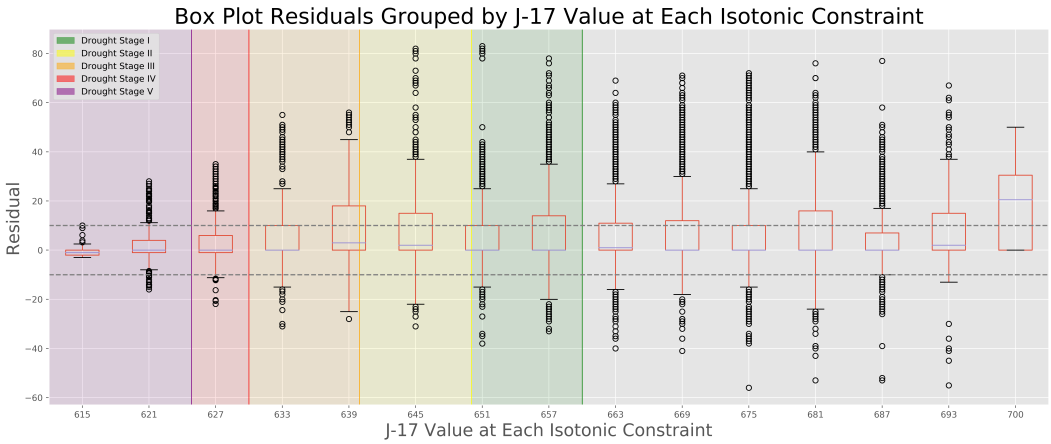


Figure 4.22: Box plots of the residuals plotted against the J-17 value at each isotonic constraint to display the distribution of predictions and the range of error after correcting for when springflow exceeds river discharge. The colors represent the EAA critical periods (see Figure 2.2). Theory-guided refinement of data science outputs was used to truncate springflow where it exceeds river discharge.

Chapter 5

Conclusions

5.1 Next Steps

Accurately predicting springflow from Comal Springs is necessary in order to protect Texas' water resources and the endangered species that reside in the spring waters. The machine learning prediction model introduced here is a faster method that requires no manual work, whereas their current method takes time each day and manual interpretation. Additionally, because the current method of springflow estimation involves uncertainties through averaging, this machine learning method may be equally as accurate. To continue this project, it would be interesting to attempt other methods to explain more of the variance in springflow. Potential for building upon this model would be to calculate a decomposition of the time series to filter out the noise. There was an attempt to incorporate pumping data for the region, but this is not available on a daily scale and it would not be useful to include single values per year. It could be of potential value to add seasonal pumping data to this model to determine its impact on springflow. Additionally, the use of bootstrap methods to further quantify model uncertainty should be explored.

The springflow prediction model presented here serves as a proof of concept that hydrologists and water resource management can examine and modify to fit their diverse needs. These methods are all available in an online repository and can therefore be incorporated into studies elsewhere. This model is specific to the regional datasets of the Edwards Aquifer and Comal Springs, but the workflow and post-processing would be similar if applied elsewhere. Applying machine learning and other data-driven methodologies in the geosciences in this fourth paradigm of scientific discovery enables scientists to focus less on the manual aspect of their research and put their efforts into interpreting the data and making decisions.

5.2 Final Remarks

Theory-based models are sufficient when working in simple, natural systems. However, they fail to fully represent relationships in complex, natural phenomena. Black-box data science models succeed in detecting underlying patterns in data, but lack the incorporation of scientific knowledge and provides little information about the underlying scientific processes. The intersection between theory-based and data science models expands our ability to benefit from the vast presence of data while incorporating domain expertise into final interpretation of model outputs. Geoscientists who use data science methods must remain a domain expert first and data scientist second. It is important to build models that align with what is possible in nature [30]. Though there are numerous challenges to applying machine learning in the geosciences [17][41], this new paradigm of scientific discovery has allowed geologists to efficiently perform analyses with big data to identify underlying patterns in natural systems and make predictions that, in the past, may have been challenging.

Bibliography

[1] Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C. Machine learning for data-driven discovery in solid earth geoscience. *Science 363*, 6433 (2019).

[2] Burnett, J. When the sky ran dry. *Texas Monthly* (Jul 2012).

[3] Ebert-Uphoff, I., Thompson, D., Demir, I., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E., and Smyth, P. A vision for the development of benchmarks to bridge geoscience and data science. *7th International Workshop on Climate Informatics* (Sep 2017).

[4] Edwards Aquifer Authority. Endangered species of the Edwards Aquifer. https://www.edwardsaquifer.org/habitat-conservation-plan/about-eahcp/covered-species/.

[5] Edwards Aquifer Authority. Water Level Monitoring. https://www.edwardsaquifer.org/science-maps/aquifer-data/water-level-monitoring/j17.

[6] Edwards Aquifer Recovery Implementation Program. Edwards Aquifer Habitat Conservation Plan. https://www.edwardsaquifer.org/habitat-conservation-plan/.

[7] Ford, D., and Williams, P. *Karst Hydrogeology and Geomorphology.* Wiley Online Library, 2007.

[8] Fuente, C. Artifical Intelligence Roadmap, year = 2019, note = https://cra.org/ccc/ai-roadmap-integrated-intelligence/.

[9] Geological Survey, U. National Water Information System data available on the World Wide Web (usgs.

[10] Gil, Y. Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery. *Data Science* (10 2017), 1–11.

[11] Gil, Y., David, C., Demir, I., Essawy, B., Fulweiler, R., Goodall, J., Karlstrom, L., Lee, H., Mills, H., Oh, J.-H., Pierce, S., Pope, A., Tzeng, M., Villamizar, S., and Yu, X. Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science 3*, 10 (2016), 388–415.

[12] Gil, Y., Pierce, S. A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., Hsu, L., Kinter, J., Knoblock, C., Krum, D., Kumar, V., Lermusiaux, P., Liu, Y., North, C., Pankratius, V., Peters, S., Plale, B., Pope, A., Ravela, S., Restrepo, J., Ridley, A., Samet, H., and Shekhar, S. Intelligent systems for geosciences: An essential research agenda. *Commun. ACM 62*, 1 (Dec. 2018), 76–84.

[13] I. Ebert-Uphoff, D.R. Thompson, I. D. Y. G. M. H. A. K. M. G. V. K. E. C.-C. P. S. Vision for the development of benchmarks to bridge geoscience and data science.

[14] Ikard, S., and Pease, E. Preferential groundwater seepage in karst terrane inferred from geoelectric measurements. *Near Surface Geophysics* (2018). https://doi.org/10.1002/nsg.12023.

[15] Kalisek, D. The grass isn't greener on the other side: Drought's effects on waterbodies, crops, livestock, energy, consumers and pocketbooks, 2011.

[16] Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering 29*, 10 (2017), 2318–2331.

[17] Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering 31*, 8 (Aug 2019), 1544–1554.

[18] Karpatne, A., Watkins, W., Read, J., and Kumar, V. Physics-guided neural networks (pgnn): An application in lake temperature modeling, 2017.

[19] Kenda, K., Senozetnik, M., Klemen, K., and Mladenić, D. Groundwater modeling with machine learning techniques: Ljubljana polje aquifer. vol. 2.

[20] Kuniansky, E. L. Precipitation, streamflow, and base flow in west-central texas, december 1974 through march 1977. Tech. rep., 1989. Report.

[21] Kuniansky, E. L., and Ardis, A. F. Hydrogeology and ground-water flow in the edwards-trinity aquifer-system, west-central, texas. Tech. rep., Reston, VA, 1997. Report.

[22] Maclay, R. W. Geology and hydrology of the Edwards Aquifer in the San Antonio area, Texas. Tech. rep., Austin, TX, 1995. Report.

[23] Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology 29*, 7 (2012), 897–910.

[24] NACE, R. L., and PLUHOWSKI, E. J. Drought of the 1950's with special reference to the midcontinent. *Geological Survey Water-Supply Paper 1804* (1965).

[25] Open Source Initiative, T. The open source definition. in: Open source initiative., 2019.

[26] Pease, E. Github Repository: ComalSpringflow (https://github.com/ecpease/ComalSpringflow), Dec. 2019.

[27] Petty, T., and Dhingra, P. Streamflow hydrology estimate using machine learning (shem). *JAWRA Journal of the American Water Resources Association 54*, 1 (2018), 55–68.

[28] P.G. George, R.E. Mace, R. P. Aquifers of Texas. *Texas Water Development Board Report 380* (2011).

[29] Puente, C. Method of estimating natural recharge to the edwards aquifer in the san antonio area, texas. Tech. rep., US Geological Survey, 1978.

[30] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature 566*, 7743 (2019), 195–204.

[31] Rutledge, A. T. Computer programs for describing the recession of ground-water discharge and for estimating mean ground-water recharge and discharge from streamflow records-update. Tech. rep., 1998. Report.

[32] Seager, R., Feldman, J., Lis, N., Ting, M., Williams, A. P., Nakamura, J., Liu, H., and Henderson, N. Whither the 100th meridian: The once and future physical and human geography of america's arid–humid divide. part ii: The meridian moves east. *Earth Interactions 22*, 5 (2018), 1–24.

[33] Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research 54*, 11 (Nov 2018), 8558–8593.

[34] Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research 54*, 11 (2018), 8558–8593.

[35] Small, T. A. Hydrogeologic sections of the Edwards Aquifer and its confining units in the San Antonio area, Texas. Tech. rep., Austin, TX, 1986. Report.

[36] Smith, D., SMith, B., Blome, C., Pierce, H., and Lambert, R. Cretaceous volcanic intrusives in the Edwards Aquifer, Texas, as identified from a high-resolution aeromagnetic survey, 2001.

[37] Survey, U. G. National water information system data available on the world wide web (usgs water data for the nation), 2016.

[38] TWDB. State Water Plan. *State Water Plan* (2017), 150.

[39] Votteler, T. H. The little fish that roared: The endangered species act, state groundwater law, and private property rights collide over the texas edwards aquifer. *Envtl. L. 28* (1998), 845.

[40] Wahl, K. L., and Wahl, T. L. Determining the flow of comal springs at new braunfels, texas. *Proceedings of Texas Water 95* (1995), 16–17.

[41] Zaidi, S. M. A., Chandola, V., Allen, M. R., Sanyal, J., Stewart, R. N., Bhaduri, B. L., and McManamay, R. A. Machine learning for energy-water nexus: challenges and opportunities. *Big Earth Data 2*, 3 (2018), 228–267.

Vita

Permanent Address: 807 W 25th Street, Apt 202, Austin, TX 78705

This thesis was typeset with LaTeX 2_ε[1] by the author.

[1] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this thesis were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.